DOCUMENT RESUME

ED 179 573 TM 009 854

AUTHOR Marchak, Nick; And Others
TITLE Assessing Communication Skills: A Review of the Literature.
INSTITUTION Alberta Dept. of Education, Edmonton. Minister's Advisory Committee on Student Achievement.
PUB DATE Mar 79
NOTE 224p.; Handwriting samples may not be legible

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Communication Skills; Educational Research; Elementary Secondary Education; Foreign Countries; Functional Illiteracy; *Listening Skills; Literacy; *Reading Ability; *Speech Communication; *Student Testing; Tests; *Writing Skills
IDENTIFIERS Canada; Great Britain

ABSTRACT

A review of the literature on assessing communications skills is presented. Communication skills are divided into five areas: reading literacy, writing literacy, functional literacy, oral communication, and listening. Within each area, a definition is given, current assessment views are listed, and descriptions of relevant research are provided. The research reported is generally limited to new assessment techniques which show potential for future applications, and to studies at the elementary through secondary levels. Activity-oriented research is emphasized. Innovative and technical research from Canada, the United States, Great Britain and other English-speaking countries is reported. Reviews are limited to research from 1970 to the present. The literature review is organized as follows: (1) assessing communication skills--an overview; (2) bibliographic essay on research for literacy assessment; (3) bibliographic essay on oral communications and listening; (4) bibliographies on each of the five skill areas, and general overview; and (5) five appendices which contain summaries of individual research projects in each of the skill areas. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Assessing Communication Skills
A Review of the Literature

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W.P. EDDY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED179573

TM 009 854

MARCH, 1979

Alberta

The Minister's Advisory Committee on Student Achievement

ASSESSING COMMUNICATION SKILLS: A REVIEW OF THE LITERATURE

A Study Conducted Under Contract to Alberta Education,
Edmonton, Alberta, Canada

BY
DR. NICK MARCHAK
SHIRLEYANNE MICHAELS
ANNE MULGREW
HELEN NEUFELD

EDMONTON PUBLIC SCHOOL DISTRICT NO. 7
EDMONTON, ALBERTA, CANADA

March 1979

ACKNOWLEDGEMENTS

In response to internal school system needs, considerable work related to assessment of Communication Skills was conducted by Edmonton Public School District personnel during 1975-76. The majority of the materials produced were subsequently incorporated into the development of this report. Acknowledgement is therefore given to the Edmonton Public School District No. 7 for its contributions.

Project Completion Date - March 31, 1977

This study was prepared by the Edmonton Public School District No. 7 under contract to Alberta Education, Edmonton, Alberta, Canada.

PREFACE

The Minister's Advisory Committee on Student Achievement (MACOSA) was established by ministerial order in October 1976 in response to growing concerns expressed by the public-at-large, government, labor, business, students and educators regarding the quality and standards of basic education in Alberta.

MACOSA commissioned a number of studies, primarily to provide basic information for a summary of current levels of achievement in Alberta and to provide baseline data for future assessment. These studies fell into three categories: (1) preliminary studies, (2) achievement studies, and (3) other studies.

This preliminary study, Assessing Communication Skills: A Review of the Literature, was designed to provide background information for developing tests of communication skills (reading, writing, listening, speaking and viewing) appropriate for use in Alberta. This initial study laid the groundwork for two subsequent MACOSA studies -- the Alberta Language Arts (Reading and Writing) Achievement Study and Test Development for Assessing Achievement in Listening and Speaking.

This report, which represents the findings and conclusions of the researchers, was presented to MACOSA as information.

EXECUTIVE SUMMARY: ASSESSING COMMUNICATION SKILLS

The assessment of communication skills demonstrated by youth has recently been the focus of wide investigations, nationally as well as internationally. Valid and comprehensive assessment of communication skills must recognize the complexity of the area as well as the interdependence among assessment, program specifications, and instructional methodologies.

In this report, communication skills have been defined to include the interrelated processes of reading, writing, speaking, and listening. The research reported for each area generally deals with new assessment techniques that show promise for future assessment applications. The major implications of research since 1970 are reported.

In the assessment of reading literacy, the majority of research studies investigated promoted the cloze technique as a more valid and reliable measurement than the traditional multiple-choice testing procedure. This technique utilizes completion exercises. The main advantage of this technqiue for assessing reading comprehension was that the examinee's task is context-dependent.

Research suggests that integrated measurement approaches hold the most promise for the assessment of writing literacy (written expression). The integrated mode would examine the extent to which written material achieves a specified purpose and desired end result as well as examine, in context, the adequacy of skills and techniques necessary to write successfully. Current assessment approaches to writing literacy stress the independent assessment of different writing components rather than the composite assessment of written material. The increased availability of computer services to assist in scoring will, for large scale studies, make the integrated assessment approach more feasible than formerly.

Research into the assessment of functional literacy, which involves both reading and writing for near-adult and adult samples, suggested that actual performance scores derived from tests measuring common daily reading and writing activities be used. The addition of assessment of functional literacy to the more academic assessment of reading and writing increases the comprehensiveness and hence, the validity of any large scale literacy assessment project.

Investigators in the area of oral communication emphasized the use of controlled but natural testing situations. There is general agreement that future assessment materials should concentrate on measuring originality in form and content rather than mechanical manipulations of form and content.

In the assessment of listening, as with oral communication, the major generalization from the research was the use of controlled but natural situations in the construction of listening tests, that is, the use of spoken English as opposed to written English. A second stress was the use of taped presentations to ensure standardization from administration to administration.

Some of the assessment trends and recommendations identified in the report could be considered as having particular promise for Alberta programs. Several of the approaches could be incorporated into the development of valid assessment instruments appropriate to the major emphasis of the Alberta Language Arts curriculum.

TABLE OF CONTENTS

	PAGE
LIST OF FIGURES	vi
ASSESSING COMMUNICATION SKILLS: AN OVERVIEW	1
Purpose of the Study	1
Areas Examined	1
Research Procedures	2
Study Findings	2
Conclusions and Recommendations	9
SUMMARY OF RESEARCH FOR LITERACY ASSESSMENT	13
Reading Literacy	13
Writing Literacy	20
Functional Literacy	24
SUMMARY OF RESEARCH FOR ORAL COMMUNICATIONS AND LISTENING	31
Oral Communication	31
Listening	35
BIBLIOGRAPHIES	43
General	45
Reading Literacy	46
Writing Literacy	52
Functional Literacy	54
Oral Communication	57
Listening	59
APPENDIX A: Reading Literacy	61
APPENDIX B: Writing Literacy	109
APPENDIX C: Functional Literacy	139
APPENDIX D: Oral Communication	161
APPENDIX E: Listening	187

LIST OF FIGURES

FIGURE		PAGE
1.	Sample Lexical Cloze Test	67
2.	Sample Passage on which Reading-Storage (RS) Test and Modified-Cloze (MC) Test are Based	73
3.	Sample Exercise from a Reading-Storage (RS) Test for the Passage in Figure 2	73
4.	Sample Exercise from a Modified-Cloze (MC) Test for the Passage in Figure 2	74
5.	A Sample of Applying the Reading-Input (RI) Procedures	79
6.	A Sample Reading-Storage (RS) Test	80
7.	The Use of Syntactic Analysis as a Vehicle for Explicity Specified Processes of Test Construction	85
8.	Underlying Structure for the Bormuth Model - Details Omitted	86
9.	Number of NAEP Writing Exercises at Different Levels of Success for Age 13 Sample	112
10.	An "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective I Social, Type One	113
11.	Scoring of "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective 1 Social, Type One	114
12.	Results of "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective I Social, Type One, for the United States Sample	114
13.	An Essay Prepared for Nine-Year-Olds on Objective III Scholastic, Type Three	115
14.	Sample Responses of the Better Papers Written by Nine-Year-Olds on the Previous Exercise	116
15.	Sample Responses of the Typical Papers Written by Nine-Year-Olds on the Previous Exercise	117
16.	Sample Responses of the Weakest Papers Written by Nine-Year-Olds on the Previous Exercise	118

LIST OF FIGURES

FIGURE		PAGE
17.	An Essay Question Prepared for Seventeen-Year-Olds on Objective IV Scholastic, Type Three	119
18.	Sample Responses of the Better Papers Written by Seventeen-Year-Olds on the Previous Exercise	120
19.	Sample Responses of the Typical Papers Written by Seventeen-Year-Olds on the Previous Exercise	121
20.	Sample Responses of the Weakest Papers Written by Seventeen-Year-Olds on the Previous Exercise	122
21.	Percentages of Different Age Samples Giving Acceptable Responses to Overlapping Tasks	124
22.	Attributes of Each Paper Examined by the Computer	127
23.	An Overview of the Analyses Performed on the Essays	128
24.	A Portion of the Statistics for the Age 13 Famous Person Exercise Papers	130
25.	Sample Functional Literacy Question from R/EAL	149
26.	Reading Occurring During General Daily Activities	157
27.	A More Detailed Sample Breakdown of the Newspapers Category Found in Figure 26	158
28.	Arrangement of a Test Booklet by Benefit Type, and General Form	159

ASSESSING COMMUNICATION SKILLS: AN OVERVIEW

Purpose of the study

Recently, the complex area of communication skills has been the subject of much concern. This concern has been illustrated by the numerous newspaper and magazine articles lamenting the decline in various communication skills such as reading, writing, and speaking. To be able to validly determine whether or not these concerns are justified, one must be able to accurately assess the status of communication skills for a given group of individuals. The purpose of this paper is to provide information which will facilitate the development of appropriate materials to assess these communication skills so that reliable and valid data on achievement levels can be generated.

Areas examined

Language usage should be assessed in a wide variety of situations with different communication demands. Successful communication must be appropriate for the situation. Communication skills include the interrelated processes of reading, writing, speaking and listening. Each communication skill involves basic skills or mechanics of the area, and appropriate activities in which the skills are applied.

For the purpose of this study, communication skills have been separated into five areas for clarity of review. These areas are reading literacy, writing literacy, functional literacy, oral communication, and listening. Within each area, a definition is given, current assessment views are listed, and descriptions of relevant research is provided. The research chosen to be reported generally deals with new assessment techniques that show promise for future assessment applications in an area. It is felt by the authors that although many of these techniques were piloted with restricted samples, minor modifications would render them appropriate for large scale testing applications.

Research reported dealt with normal samples, for whom English was generally the native tongue. The majority of the studies reported dealt with kindergarten to secondary school-aged children. Several studies with first and second year university students were reported when the results were felt to be generalizable to senior high school samples. Activity-oriented research was reported whenever possible as opposed to pure research as the results seemed to have more applicability to the purpose of the current study.

With respect to each of the five areas, efforts were not concentrated toward presenting materials on traditional commercial testing instruments for two reasons. First, a review of current research indicated that many traditional approaches have been shown

to be invalid, or lacking in the measurement of the complex areas of communication skills. Within each area, examples of major problems with respect to lack of validity have been discussed. If it was desired to assess communication skills to the limited extent possible by these commercial tests, many ready references such as Buros, and test publisher catalogues exist with details as to the instrument's suitability and statistical characteristics.

Second, this study does not concentrate on large sampling or system testing programs. It was felt that information on the administrative timetabling of existing commercial language arts tests at the provincial or state level would not provide the kind of background desired for this study. However, the most important projects from an innovative and technical viewpoint found in Canada, United States, Great Britain, and other English-speaking countries were discussed.

Research procedures

Investigations in communication skills were limited as indicated above, and only the assessment portion of communication skills was reviewed in the literature. To confine the view to current approaches, the literature from 1970 to the present was examined. A number of sources were used to attempt as comprehensive a review of the literature as possible. Main sources that were used include the Education Index, a major source of educational periodicals; Educational Resources Information Center (ERIC), a United States Office of Education clearing house which collects and disseminates reports of current educational research; Research in Education (RIE), a listing of research abstracts; the Canadian Education Index, a major source of Canadian educational periodicals; current books dealing with the assessment of communication skills; and original source information where appropriate.

Using the above sources, pertinent articles were collected and summarized to provide source data for this report. The summary of each study includes purpose, instrumentation and methodology, analysis procedures, major outcomes and interpretations wherever possible. These summaries have been included in the appendices.

Study findings

Communication skills, as they are displayed in the Alberta Elementary Language Arts Handbook (1973), are divided into two broad areas, the receiving component and the expressive component. These components are further subdivided. The receiving component includes listening and reading while the expressive component involves speaking and writing. This study then, has used these divisions, but has also incorporated a fifth category involving functional reading and writing as it applies to adult or near adult populations.

In current literature, three of these components are commonly considered together as different aspects of formal literacy. These components are reading literacy, writing literacy, and functional literacy. Literacy in this sense means the ability to understand what one reads and to write what one understands. The main difference between the functional component and the other two is the level of competence required. The function of literacy assessment then is to determine whether or not a person possesses these abilities.

Reading literacy may be thought of as being able to respond appropriately to written language. This response may take different forms as the reader strives to reconstruct the meaning of the materials he is reading. Going one step beyond this definition, a good reader understands not only the meaning of a passage but its related meaning as well. This includes all the reader knows that enriches or illumines the literal meaning. Thus, the range of behaviors one can expect of a person considered literate in reading is complex. Because of this, undertaking the assessment of these behaviors becomes a complicated task.

Writing literacy which is expressive in nature involves again an extremely complex area. It includes the possession of both the skills and techniques required for successful writing, as well as the actual writing activities themselves. Valid assessment with respect to writing cannot limit itself to either one or the other aspect. The most effective assessment must involve the integration of both.

Functional literacy can be defined as having the basic reading and writing skills necessary to participate in a mechanized, automated society. Without these minimal skills, an individual cannot successfully function in an independent mode. Since this aspect of literacy stresses functional skills, the assessment thrust has concentrated on materials that would typically be found in real life situations.

Oral communication can be defined as a spoken unique communication that conveys an idea, feeling, or experience to others. To achieve this communication, the speaker must bear in mind the effect to be achieved, the nature of the audience he is attempting to communicate with, the appropriate form of communication to be used, and the material to be communicated. Again, oral communication can be divided into basic speaking skills and speaking activities. Valid assessment must consider both of these areas.

Listening can be defined as the act of receiving oral language. This receiving skill is essential in today's world where the spoken word is the most powerful medium of communication. Responsible listening demands more than attention to the extrinsic conditions which make auditing possible. After the stimulus has been received, it becomes the active function of the listener to

interpret, evaluate, and respond to the message, meaning, and mood of the sender. A valid assessment program would necessarily reflect these different listening components.

Having defined the five areas, major assessment findings with respect to each area will now be discussed. The findings that have been reported tend to reflect the current assessment stress in each area. It is interesting that research in different countries tended to respond to common concerns in a similar manner. Thus, work initiated in one country was being independently undertaken in other countries, the findings in most instances supporting each other.

In considering the five components of communication skills, reading has been given the most attention from an assessment viewpoint. With the current emphasis on both reading and written literacy, methods other than the traditional techniques for assessing reading literacy have been investigated. Traditionally, and in large-scale studies as late as 1973, the passage-multiple-choice type test was used as the typical assessment mode. Although numerous in-depth studies have been undertaken in many countries to determine if reading literacy is declining or improving, only conflicting reports have resulted. For example, Farr (1974) found after an extensive review of the literature and a survey of one hundred school systems that it is extremely difficult for anyone interested in evaluating trends in literacy to obtain enough adequate data to determine reading competency trends.

Several national or international organizations have attempted to answer the question of whether reading literacy is declining by conducting large scale surveys of reading. Researchers such as the International Association for the Evaluation of Educational Achievement (IEA), Inner London Education Authority (ILEA), and the National Assessment of Educational Progress (NAEP) have undertaken traditional measures of reading over the past few years in an attempt to answer this question. Interpretations of the results of these studies have been disputed by authorities in the field, and in fact appear to have left more questions unanswered than answered. Of these three projects, that undertaken by NAEP shows the most promise for future studies of this type.

Much of the controversy associated with the above projects has centered on the type of testing instrument used. Traditional measures of reading comprehension are readily available, but they appear to measure only word recognition and literal comprehension to some degree. Major limitations of the traditional type of standardized reading test have brought about investigations of possible alternative test forms. The most promising recent work to overcome the problems posed by traditional test types has been the development of the cloze technique. The basic cloze technique involves deleting every nth word from a prose passage, and asking

the respondent to complete the passage. Although not originated by Bormuth, the cloze technique has been extensively researched and developed as a measurement technique by him (1969, 1971, 1972, 1973-74). His factor validity studies in 1969 indicated that these completely objectively-developed tests measure reading comprehension skills similar to those measured in standardized reading tests. At the same time, cloze tests have the added feature of being context-dependent, which has been found not to exist commonly in traditional tests (Tuinman, 1972, 1973-74; Pyrczak, 1974, 1975, 1976; and Scherich and Hanna, 1976).

Numerous researchers have further investigated the use of the cloze test format to measure reading comprehension. Some of the more interesting applications of the basic cloze technique are listed below. Carver (1970, 1971a and b, 1974a and b, 1975a, b, and c) has done extensive work in the area by developing two test modes that were variations of the basic cloze technique, reading-input (RI) and reading-storage (RS). The RI test was found to be the most useful for measuring readability of a passage while the RS test was most appropriate for measuring reading comprehension. Balyeat and Norman (1975) developed an interesting combination of cloze and multiple-choice format using reading selections obtained from students writing or dictation. Applications of this technique have been discussed by Robinson (1973) and James (1975). Neville and Pugh (1975 and 1976-77) varied the approach to cloze tasks by modifying an existing test to both a listening test and a restricted-display version. O'Reilly (1975) has incorporated cloze passage test development with computer-assisted test assembly. More recently, Olson (1976), has built criterion referenced cloze reading tests. Many other investigators have been active in promoting the use of the cloze technique as a more valid and reliable way of measuring reading comprehension.

An intensive review of the literature has indicated that little work has been done in the area of assessment of writing literacy. Bullock (1974) summarized the status of assessment of writing literacy in Great Britain by saying no national attempt had been made to monitor standards of achievement in writing. He suggested that future emphasis should not concentrate on the isolated measurement of writing skills, but rather on the assessment of writing activities and where possible, the integration of activities with skills.

Current investigations in North America have been undertaken with this type of emphasis evident in the assessment exercise used. The major agency working in this area has been National Assessment of Educational Progress (NAEP). Their project has been ongoing for over seven years indicating the recent importance the United States places on integrated assessment of writing skills and activities. Instruments designed to assess the status of writing activities throughout the United States were administered using stratified sampling techniques. This assessment included four types of analyses of writing compositions as well as a separate analysis for

successful presence of primary trait. Since all grading was done by readers specifically trained in scoring techniques involved, inter-judge reliability was high and comparability of grading was more assured that would be expected in traditional approaches.

The four types of analyses included holistic scoring, a mechanical error count, a characterization analysis, and a computer analysis of writing skills. Holistic scoring was an innovative way of looking at initial essay scoring. Papers were ranked for overall quality on an 8-point scale and thus assigned scores reflected quality relative to other papers rather than quality determined by a set criteria. This permitted further analyses to assist in describing characteristics of each of the eight categories. A second analysis was mechanical error count where predefined errors were tabulated for given quality level papers. This provided an opportunity to quantitatively describe basic writing skills, without fragmentation in the assessment approach. This information was further enhanced by the characterization analysis. Here, correct and incorrect aspects of the papers were discussed with respect to the nine areas reviewed in the error counts to provide interpretive information with respect to the characteristics of each of the eight levels of writing. A final analysis included a computer computation of six qualities of writing. These qualities were fluency, spelling, diction, sentence structure, punctuation, and paragraph development.

The Primary Trait System (1975) is the most recent analysis technique introduced by NAEP. The essential task of these exercises was to determine if a writer could accomplish a specified purpose. Secondary traits were also included in some writing exercises, and were assessed in the same manner. The innovative aspect of all these analyses is that component scores are awarded rather than composite or total scores.

The study of functional literacy is relatively recent. Many of the testing tasks in this area are novel and would not be considered part of a formal traditional assessment. Adult illiteracy has recently been designated a major target area of the Right to Read Program in the United States. The extent of their commitment to reduce adult illiteracy is reflected in a national goal of eliminating functional illiteracy by 1980 among 90% of the population over 16 years of age. The desire to determine the efficiency of programs to reduce illiteracy has led in turn to a need for instruments that measure functional literacy. Thomas (1976) has surveyed cross-Canada activities in literacy under World Literacy of Canada auspices. He found that 37 percent of the out-of-school adult population over 15 years of age had no more than elementary schooling.

Since 1970, many organizations and individuals have attempted to develop operational definitions for functional literacy, approaching it from either one of two viewpoints. Either functional literacy has been considered as the basic reading and writing skills needed to participate in society at a survival level or it has been thought

to represent an individual's status on essential real-life skills. In either case, the practical rather than the academic has been stressed.

Testing materials for minimal competencies have been developed by the National Reading Council in New York (1970), by Sticht et al (1973) for American Military selection, the Maryland Department of Education (1972), Lichtman (1973), and by Murphy of the Educational Testing Service(1973). An essential characteristic of all these tests was the use of application forms, schedules, simple directions, signs and other simulated real-life tasks.

Other studies have concentrated on the development of assessment instruments to determine levels of proficiency of high school graduates on practical materials which both the school and community would expect them to read and use. Three projects that have recently set these baseline levels are the criterion-referenced reading test of McDonald and Moorman (1974), the development of a cloze test by Hansen and Hesse (1974), and the NAEP Mini-Assessment of Functional Literacy under the Right to Read Program (1975). The exercises used in these types of tests, although still practical in emphasis, tended to have a higher reading level and draw from a wider range of sources. Adult literacy testing is still a developing field marked by assessment instruments of varying quality.

Oral communication has been considered with different degrees of importance in the formal educational structure. Studies of oral language, particularly spontaneous utterances have been hampered by the complexity of analyses procedures required. A second concern of any oral language measurement is the degree of subjectivity involved. A third concern involves the time availability in our present teaching structure, and the lack of qualified classroom personnel to undertake extensive teaching or assessment of oral communication. For valid oral communication assessment to take place, these concerns must be handled, and the assessment procedures must comprehensively include both basic oral skills and oral activities. Research at this time has generally not reached this desired state, but has tended to concentrate on particular aspects of either oral skills or oral activities. Thus, the reporting of the findings is organized into these two areas.

The assessment of basic oral skills has included work by Carrow (1974) where a sufficient quantity of predetermined sentences were repeated by children to simulate spontaneous speech patterns; by Graham (1971) where the use of pictures and maximum cueing encouraged subjects to produce particular verbal structures within their competence; and by more practical researchers such as Cayden (1975) and Dobbyn (1976) who structured controlled natural group situations for the assessment of basic oral skills.

More research has been carried out in the oral activities section. Studies to assess student ability to describe and inter-

communicate with peers were designed by Williams et al (1970), Rose et al (1973), and Melear (1973). Other investigations have been concerned with the effectiveness of students' oral presentations, (Williamson, 1969; Skull and Williamson, 1969; Tucker, 1970; Robinson, 1971), or objective assessment of group oral conversation (Tucker, 1970; Folland and Robertson, 1976). Stauffer and Pikulski (1974) undertook further work that showed the more practice young children have in using oral skills, the more proficient they become; while Potter and Strother (1975) showed that counter-productive responses in oral communication could be significantly decreased through behavior modification techniques. These final two procedures both show promise for objective continuous measurement of both student progress in oral language and the effectiveness of a speech development program.

Educators within the field of reading have long emphasized the importance of listening with respect to reading, and have contended that a measure of listening comprehension should be predictive of reading achievement. As a recent result of this, there has been an increasing emphasis on the place of listening in education. Wilkinson (1970) states that curriculum construction and assessment should be thought of both in terms of written literacy and oracy which includes both speaking and listening. His work in the area has attempted to promote this viewpoint.

Current viewpoints in the area of listening assessment have established certain general criteria for valid testing. Bullock (1975) stated that listening tests should be representative of the listening skills the individual needs for his daily varied activities. Wilkinson et al (1974) advised that subjects should have the opportunity to respond to many pieces of English before their listening ability is judged. They also indicated that the best testing materials are taken in context rather than as isolated items. Finally, they indicated that these contextual stimulus selections for listening tests should be presented from standardized tapes. Any measurement of listening then must attend to these guidelines for quality listening assessment.

Prior to 1950, listening was not assessed as a separate component of an English program. One of the first comprehension tests was developed in 1953 by Brown and Carlson. This initial attempt at measuring listening skills, although innovative for its time, was limited by the fragmentary nature of many of its test questions and by the fact that the test was composed of written language read aloud.

A question of interest to researchers has been whether or not listening skills can be taught. Studies by Hackett (1970), Penfield and Marascuilo (1972), Palnatier and McNinch (1972), and Rosner (1974) found that while listening skills could be taught, the most effective learning took place at the early grades, while no effective gains were observed for high school students. As a

by-product of these investigations, a number of interesting approaches to assessing listening were developed.

A new dimension in the assessment of speech-sound discrimination was developed by Goldman et al (1971). Their instrument evaluated speech-sound discrimination ability in a novel way. While this instrument is finding many applications, and the general consensus is that the test has merit, further research remains to answer all questions concerning its validity and reliability. Listening skills have also been studied on a more theoretical level by investigators such as Mecham (1971), Rosner and Simon (1971) and Schwartz and Goldman (1974).

Other researchers have concentrated on the measurement of listening activities (Nesbit, 1968; Butler, 1970; Wollner, 1974). Some of these studies have been concerned with demonstrating that a predictive relationship exists between listening skills and reading achievement. Further work on the assessment of comprehension aspects of listening ability by Bowdidge (1967) and Kennedy and Wiener (1973) showed that the cloze procedure could be used to assess listening as well as reading.

The use of natural settings in any form of listening assessment has been the major thrust of work undertaken in Great Britain. In the more informal types of assessment appropriate for the classroom setting, it is not difficult to incorporate the criteria of natural setting. More problems are encountered when one moves to a more formal type of assessment. Wilkinson et al (1974) have dealt with these problems very effectively in the construction of three Listening Comprehension Assessment batteries for children aged eleven to eighteen. These batteries appear to have incorporated as much as possible the general criteria specified for valid testing of listening.

Conclusions and recommendations

This study has been concerned with five aspects of communication skills. Operational definitions for each aspect have been provided and relevant research on sound assessment practices have been reported for each area. By summarizing these investigations, general conclusions have been generated with respect to each area. The conclusions to be reported here contain generalizations with respect to the most promising kinds of assessment practices, and suggest profitable directions in which further assessment research or application might be considered.

Until recently, most assessments of reading literacy have utilized the traditional multiple-choice type of test question. More recent practices are moving away from this type of standardized test for assessing reading. The cloze technique is being promoted as a more valid and reliable means of assessing reading literacy.

Tests developed using this technique have been shown to measure skills similar to those measured in the more traditional tests while at the same time, ensuring that the examinee's task is context-dependent. Modifications to the cloze technique are allowing additional more complex skills to be measured. Increased computer assistance could greatly facilitate advances in this area.

Future valid assessments of writing will of necessity be more comprehensive than in the past. integrated measurement approaches will be used rather than the examination of isolated skills. However, this does not mean that a composite assessment score will be calculated for a piece of writing but rather that the student will perceive the exercise as an integrated whole. The assessment of the exercise will provide information on different writing components as well as descriptive info mation on the whole piece of writing. The increased availability of mechanical aids to assist in the assessment makes this type of scoring more feasible than formerly for large scale studies. At present, the only existing materials available for use in this way for the assessment of writing literacy are those developed by NAEP. Some attempts are being made in Great Britain to develop similar materials. In the opinion of experts in the area, standardized commercial component tests such as Canadian Tests of Basic Skills and the Cooperative Tests by ETS do not measure the essential qualities of writing literacy.

There appears to be a general consensus among the researchers who have investigated assessment of functional literacy over the past six years that any index used to measure literacy must be more than simply a grade level passed in school. Rather, such an index must be based on an actual performance score of the individual. Tests used to obtain this index should be structured using a criterion-referenced model. Available tests that appear to show promise for assessing functional literacy include the Functional Literacy Test, R/EAL, MRPA, MAFL, or the tests developed by ETS in their adult Functional Reading Study. All of these employ forms, signs, letters, recipes, instructions and other daily encountered tasks. With the current public interest into the practical aspect of literacy assessment, it would appear that the addition of assessment of this area to the more academic assessment of reading and writing would increase the comprehensiveness and hence the validity of any large scale literacy assessment project.

Although little attention has been paid to the assessment of oral communication as part of the language arts program in the past, current emphasis on this area has resulted in small-scale assessment investigations being undertaken in both the areas of basic oral skills and oral activities. Investigators emphasize the importance of controlled but natural testing situations. The lack of standardized commercial instruments for measuring oral language competencies is evidence of the special problems associated with this type of assessment. To this point, the most successful

testing has taken place on an individual or small group basis. Experts agree that future assessment materials should concentrate on measuring originality in form and content rather than mechanical manipulations.

The responsible practicing of the skill of listening is becoming an increasingly important part of communication abilities. As a result, assessment investigations have been undertaken in both the areas of basic listening skills and listening activities. One of the most important generalizations was the use of controlled but natural materials in the construction of listening tests - use of spoken English as opposed to written English. Continuing this generalization, it has been shown that test results are more valid when materials used are placed in context. The use of taped presentations to ensure standardization from administration to administration was stressed. Few valid commercial tests currently exist for the assessment of listening. Two major exceptions are Goldman-Fristoe-Woodcock Test of Auditory Discrimination (1970), and the three Listening Comprehension Assessment Batteries developed in Great Britain by Wilkinson et al (1974). The most promise for formal assessment in listening lies in the further development of test batteries such as these authors have illustrated.

This section has attempted to point out the major recommendations with respect to assessment of the five areas of communication skills. Further study and investigations will permit the enlarging and substantiation of these viewpoints, and permit the development of valid assessment instruments appropriate to the major emphasis of the Alberta Language Arts curriculum.

SUMMARY OF RESEARCH FOR LITERACY ASSESSMENT

Literacy may be defined as the ability to understand what one reads and to write what one understands. Critical literacy should provide the tools to comprehend one's experience in order to change it. This implies not only comprehension of material but also careful analysis and judicious evaluation of the information received. The function of literacy assessment then is to determine whether or not a person possesses these abilities. For purposes of measurement, literacy may be divided into three major components: reading literacy, writing literacy, and functional literacy.

Reading literacy assessment

Definition. Reading literacy may be thought of as the ability to respond appropriately to written language. The purpose of reading is the reconstruction of meaning. Meaning is not in print, but the author begins with meaning when he writes. Somehow, the reader strives to reconstruct this meaning as he reads. Going one step beyond this definition, a good reader understands not only the meaning of a passage, but its related meaning as well, which includes all the reader knows that enriches or illumines the literal meaning. Here, the reader is required to engage in critical and creative thinking in order to relate what he reads to what he already knows.

Deciding what types of behaviors one ought to expect of a literate person presents one type of problem, but deciding how those behaviors should be observed and assessed presents complications of a different nature. It must be recognized that what we are forced to observe in assessing reading literacy is not the processes that we really want to observe but merely overt behaviors that we take as indicating the presence or absence of these processes. It must also be accepted that due to the incomplete nature of the theories regarding the processes underlying literacy behaviors, it is not possible at the present time to have operational procedures for testing those processes in a way that each process is individually identifiable. However, techniques are being developed which would allow testing of at least the decoding and various comprehension processes involved in literacy behaviors even though each individual process cannot be isolated. A third consideration involves the specification of a level of performance a person must exhibit on a test before he may be regarded as literate. Considerable ingenuity and resourcefulness are required to establish this criterion level for reading literacy.

Views on reading literacy. Is there a decline in literacy? What are current reading standards? Are they changing? These questions have been debated by numerous researchers. Among these researchers are Ledbetter and Daniels (1976) and Farr (1974) from the United States; Badcock (1974), Lethborg et al (1974), Raig (1974),

and the Educational Magazine from Australia; and Welton (1974) from Canada. These writers have discussed both sides of the debate of declining literacy. Burke (1975), after critically reviewing six studies of the reading standards of British children conducted between 1966 and 1972, found no justification for the belief that standards are declining. Bookbinder (1973) and Amberden (1974) also questioned the reliability of reports of declining standards in Britain and suggested that misleading information and gross extrapolations may have led to false conclusions regarding declining standards. Farr (1974), following extensive review of the literature and the surveying of one hundred school systems in the United States, concluded that it is extremely difficult for anyone interested in evaluating trends in literacy to obtain adequate data necessary for determining reading competency trends.

Several authors have published collections of recent reading research. Goodacre (1974, 1975) listed research conducted in Britain in 1973 and 1974 and Weintraub et al (1974-75, 1975-76) listed American reading investigations conducted in 1973 and 1974. Sherwood (1976) provided a historical perspective of the entire issue of evaluation in Great Britain over the past century. Research contained within these listings illustrated the conflicting views with respect to standards of reading literacy found in the English-speaking world today.

Assessment of reading literacy. Several national or international organizations have attempted to answer the question of whether reading literacy is declining by conducting large scale surveys in reading. These researchers have generally utilized the more traditional measures of reading such as the passage followed by multiple-choice question type or the multiple-choice sentence completion type. Examples of large scale surveys are noted below.

The International Association for the Evaluation of Educational Achievement (IEA) in Stockholm (Schwille and Marchall 1974) surveyed both Reading Comprehension and Literature as part of a 6-subject international survey on achievement. Tests developed and used were of the passage followed by multiple-choice questions variety. Results from the 1970-73 survey were reported by Thorndike (1973). Several of the major findings reported were that (1) achievement in developing countries fell far below that of countries having a relatively high level of economic development, (2) an appreciable prediction of the reading achievement of individual students in developed countries was provided by information about their home and family backgrounds, and (3) there appeared to be little evidence of the impact of the school or of specific factors on the progress of students in reading. Downing and Dalrymple-Alford (1974-75) provided a methodological critique of Thorndike's report. The report was criticized on the basis of unclear reporting, lack of logical framework, methodological weaknesses and over-generalized conclusions. A similar review by McCullough et al (1974) also noted several major problems of the Thorndike report. Both the Downing and

McCullough reviews concluded that the Thorndike report is best considered a pioneer study in the area rather than a basis upon which to formulate conclusions regarding the status of literacy.

Inner London Education Authority (ILEA) (1969) conducted a Literacy Survey with follow-ups in 1971 and 1974. Tests used in these surveys were developed by National Foundation for Educational Research in England and Wales (NFER). These tests were of the multiple-choice sentence completion type or the passage followed by multiple-choice items type. The 1968 survey indicated that eight-year-olds in the sample were reading about six months behind the national standardization group. The 1971 follow-up showed a widening gap between chronological and reading age even though the mean standard score remained the same. However, the same sample in 1973-74 had somewhat improved their position. Varlaam (1973) for the ILEA reported that a sizable portion of the children tested on the Rutter Behavior Scale in the 1971 follow-up were found to be abnormally anti-social or neurotic or both. This abnormal behavior was found to be frequently accompanied by backwardness in reading, verbal reasoning, English and other subjects. Burke (1975) questioned the conclusion of the ILEA survey that standards were declining and suggested that since this decision was based on comparisons of scores with norms of one test only, caution should be exercised in formulating generalized conclusions regarding the status of literacy in Britain. Amberden (1974) suggested that the NFER tests used in major surveys in Great Britain were out-of-date and subject to misinterpretation. She further stated that trends in reading standards based on NFER test may have been completely without basis. This stance was consistent with widespread controversy regarding changing literacy standards as reported by NFER studies.

National Assessment of Educational Progress (NAEP) in Denver has conducted cross-country studies over the past seven years to assess student performance on specific reading objectives rather than reading comprehension in general. Numerous states have patterned their statewide assessments after NAEP national survey according to Forbes (1977). For example, he indicated that the Connecticut assessment showed deficiencies in the reading skills of youth from large cities while students from small communities were found to be performing above national achievement levels. Alvord and Brittingham (1974) have identified problems associated with the use of difficulty levels assigned by NAEP to their exercises. Caution was suggested for use of these exercises in establishing criterion levels for reading objectives. Shafer (1971) considered the NAEP assessment as a means to communicate more effectively with laymen as well as educators.

Traditional measures of reading comprehension such as those used in the above studies are readily available and do appear to measure word recognition and comprehension to some degree. However, research has indicated a major deficiency of most currently available reading comprehension tests. It is often possible for students to answer many of the items correctly beyond the level of chance

without reading the passages (Tuinman 1972, 1973-74; Pyrczak 1974, 1975, 1976; and Scherich and Hanna 1976). Such seemed to be the case for most standardized tests on the market. Even in test items specifically designed to measure vocabulary in context, most items were not context-dependent. Content validity may be seriously in question for a test on which students can successfully complete items measuring "meaning in context" without reading the context. The Davis Reading Test (Pyrczak 1976) was an exception when tested in studies of context-dependency.

Such major limitations of the traditional type of standardized reading test have caused individuals to investigate possible alternative test forms. Carver (1973) reviewed the influence of E. L. Thorndike's testing technique (presenting paragraphs with questions based on them) on many currently available standardized reading tests. Carver contended that these questions may in fact, not indicate progress in reading as much as progress in reasoning. He concluded that what is needed are tests which actually measure absolute levels of the ability to read sentences that make up paragraphs rather than the ability to answer reasoning-type questions on paragraphs. Several large scale users such as Educational Testing Service, Princeton, and National Foundation for Educational Research in England and Wales have also been considering new testing alternatives. Spalding (1974) reported a recent major change in NFER policy regarding the type of test to be used. Plans called for the development of item banks of testing materials which would be passage-dependent. Such questions were to be available for use for criterion-referenced testing by classroom teachers as well as by research personnel.

Current research in the area reflects a major trend away from the traditional types of reading comprehension tests. The procedure which is frequently being utilized in the United States, Great Britain, and Australia involves the application of one or several different "cloze" techniques. The basic cloze technique involves deleting every n^{th} word from a prose passage and asking the respondent to complete the passage. Numerous variations of this technique have been employed by a variety of researchers. An overall review of the cloze procedure was presented by Bickley et al (1970); Robinson (1972) has provided an annotated bibliography of the cloze procedure to that date. Items for cloze tests were produced from prose passages through the use of an algorithm, thus ensuring complete objectivity in item construction. Recent researchers (Finn 1975 and Carver 1975c) further refined these algorithms to ensure that independent item writers would generate not only the same but legitimate achievement test questions from any passage. This would ensure complete replicability of any test development study using the cloze procedure.

Although not originated by Bormuth, the cloze technique has been extensively researched and developed as a measurement technique by him (1969, 1971, 1972, 1973-74). His factor validity

studies in 1969 indicated that these completely objectively developed tests measure reading comprehension skills similar to those measured in standardized reading tests. Research that supports this generalization was conducted by Neville and Pugh (1974), and by Balyeat and Norman (1975).

Bormuth provided an extensive operational model for utilizing cloze techniques and related procedures in the measurement of reading comprehension. The first component included in the model was a taxonomy of literacy behaviors consisting of the following skills: decoding, literal comprehension, inferential comprehension, critical reading, aesthetic appreciation, reading flexibility, and study skills. Also provided was a definition of the criteria used for selection of reading tasks, and a discussion of the measurement of the literacy behaviors. Bormuth outlined a mathematical model for identifying criterion levels of performance that can be rationally defined as being the best level of performance for a particular instructional task. Seven basic steps were required in order to implement the model. A score of thirty-five per cent on cloze tests was suggested as the criterion score for minimal reading literacy. Bormuth developed the initial algorithm for constructing cloze items and established procedures for determining the difficulty level of cloze passages to measure his defined reading behaviors.

Numerous researchers have further investigated the use of the cloze test format as a means of measuring reading comprehension. Variations of the technique are many. Some of the more interesting applications of the basic cloze technique are discussed below. Carver (1970, 1971a and b, 1974a and b, 1975a, b, and c) has done extensive work in the area of using a modified cloze technique to measure reading comprehension. In the 1971a paper, the reading program of an individual was analyzed into two components: reading-input (RI) and reading-storage (RS). Rules or algorithms for developing these two types of tests had been previously developed by him (1970). An RI item appeared similar to the original passage except that alternative words were added to the correct item words. The task for the subject was to select the correct word for each item. An actual reading-storage test was developed and evaluated (1971b). The RS test was actually a type of cloze procedure in which every other word was deleted except for the initial letter. The remaining words and blanks were separated into 10 lines with 5 blanks per line. One of the five initial letters was replaced by a wrong letter. The subject was expected to read the passage, and then recognize and circle the wrong letter on each line without referring back to the passage. Carver (1974b) further evaluated the RS test by comparing it to his modified version of the cloze test. The RS test seemed to be more reliable and equally as valid as the modified cloze test in measuring gain in understanding due to reading, that is, reading comprehension. Carver (1975a) then compared the RS test to a paraphrase test (Anderson 1972) to determine the sensitivity of each to the gain in comprehension that resulted from reading prose.

The completely objectively developed RS test appeared to be as sensitive to comprehension gain as was the subjectively developed paraphrase test. Carver (1975b) presented revised algorithms for developing both the RS and RI tests. These revised algorithms were easier to use and provided more valid test questions.

Carver further discussed the RI and RS tests in comparison to multiple-choice and regular cloze items with respect to several possible applications. The RI test appeared to be the best measure of readability of the four tests, that is, the best indicator of the difficulty level of a prose passage. Carver further concluded that the RI test appeared to be the best measure to induce prose learning and showed potential as a method to improve reading skill. The RS test, being passage-dependent, was most appropriate for measuring gain due to the primary effect of reading, that is, reading comprehension. A further investigation of the RS test as a measure of gain was made by Carver (1975c) in which gains were related to passage difficulty, reader ability, and to the judgments of the reader as to how well they understood the passage. Results indicated that scores between reading conditions where the passage is read ahead of time and is not read ahead differed more for easy material than for difficult material. These gain differences were also greater for high ability subjects as compared to low ability individuals. Carver concluded that the RS test could be used to indicate the degree to which a passage was understood. Since the RS test appeared to be sensitive to gain due to the reading as long as the difficulty level of the material was approximately equal to the reading ability level of the individual, this type of test would likely be a reliable measure of reading comprehension.

Other researchers have concentrated only on the application of a cloze technique to determine readability. Peterson et al (1973) have also recommended the use of a cloze procedure as a means of measuring the instructional level of material. Klare (1974-75) reviewed readability formulas and predictive devices used since 1960. He indicated that the cloze procedure was particularly adaptable for measuring readability in a number of languages. Klare also stressed that while readability formulas provide good indices of difficulty, they do not indicate causes of difficulty. The use of a cloze technique as a readability measure has been shown to be more reliable than traditional readability formulas for computing a readability index.

An interesting combination of cloze and multiple-choice format was utilized by Balyeat and Norman (1975). In the lexical cloze multiple-choice technique, nouns, main verbs and adjectives of a passage were identified. Every fifth word which was one of these parts of speech was deleted. These five words were paired with distractor words (1) from the same grammatical class, (2) from the grade level at which the word first frequently appears, and (3) matching the type of word, that is, general or core word, or a specific content word. These ten words were listed across the page

in random order, just after the line of copy in which the fifth deletion appeared. The Language Experience Approach was integrated with cloze technique (LEA-cloze) where selections for passages were obtained from student writing or dictation. Balyeat and Norman concluded that the LEA-cloze had value in either predicting or indicating reading levels of disadvantaged rural children. This technique has proved to be a definite motivational aid in applications for both the disabled reader and special education students (James, 1975). The LEA-cloze has also been successfully used with the disabled adult reader (Robinson, 1973) and may be used as a supplement to cloze passages made from selections not directly student originated.

Several other interesting applications of the cloze technique have been made by various researchers. Kennedy (1974) has suggested using the cloze technique to teach comprehension skills; Aaron (1975) used the technique to develop a group Informal Reading Inventory. Neville and Puch (1974 and 1976-77) varied the approach to cloze tasks by modifying an Australian-developed and British-standardized cloze test (Gapadol Comprehension Test) to both a listening test (by application to an audio tape) and a restricted-display version. Page (1975) investigated the relationship between conventional cloze test performance and a post-oral reading cloze test where the cloze items were written after reading the original passages aloud. It appeared that conventional cloze test scores could be predicted reasonably well from post-oral reading cloze test scores. Since conventional cloze tests have been shown to predict literal comprehension as measured by multiple-choice tests (Bormuth, 1969), the predictive potential of the post-oral reading test seemed to represent a valuable link between oral reading and comprehension performance. Recently, computer-assisted test assembly has been incorporated with cloze passage test development (O'Reilly, 1975). One of the first large-scale assessment programs using a version of the cloze test technique has been successfully completed in Texas (Olson, 1976) where two cloze criterion-referenced reading tests were developed, the Paragraph Reading Test and the Survey of Reading Skills for use with students from elementary through secondary school.

Overview. The public outcry of declining standards in reading literacy has caused many debates and investigations into the issue. Both armchair discussions and vigorous investigations have resulted. Large and small scale surveys have been conducted, reviewed and criticized as each tried to answer the questions. Until recently most assessments of reading literacy have utilized the traditional multiple-choice type of test question. Latest trends in research, however, have indicated a move away from this type of standardized test for assessing reading. One very basic reason for this was the lack of context-dependence of most available standardized reading tests.

Many researchers in various countries are now investigating the merits of the original objectively developed cloze technique or a variation of this technique as an alternate measure of reading comprehension. Significant improvements in algorithms used to develop cloze test items now ensure a high degree of replicability of tests developed in this manner. These tests also seem to be measuring skills similar to those measured in the more traditional tests. However, modifications to the technique are allowing additional more complex skills to be measured. Some versions of the cloze tests appear to be excellent indicators of passage difficulty. Such tests could thus be used instead of a readability formula to determine the appropriateness of prose passages. This same type of test could also be used as an instructional tool to induce prose learning and shows good potential as a method to improve reading comprehension. When the Language Experience Approach was integrated with cloze, the technique proved to be a strong motivational device for various groups of disabled readers. A major adaptation of the cloze technique to a listening test also has potential for numerous assessment and instructional applications.

It thus appears that while the more traditional multiple-choice test does have a limited place in very large scale survey testing situations, the investigator who is interested in using more valid measurement techniques, particularly for the assessment of reading comprehension, will choose to use the cloze technique. Suitable applications of the cloze technique appear to be many and extensive. Use of the cloze technique or a modification thereof would provide a valuable tool for the assessment of reading literacy.

Writing literacy assessment

Definition. Writing literacy, although difficult to operationally define, includes two main components. The first involves the skills and techniques necessary to write successfully. This component often has included such topics as sentence structure, writing formats, capitalization, punctuation, grammar usage and sentence construction techniques. The second component involves the actual writing activities themselves. Writing activities have been more concerned with purpose and desired end result, thus different forms and intended audiences were stressed as students wrote letters, reports, summaries, stories, poems, plays, essays, and so on.

Foley (1971) indicated that complete assessment of writing literacy would include measurement of the development of the selection of ideas according to relevance, logic and subordination; organization of compositions; style; mechanics; and choice of words. These should be assessed in the context of an integrated piece of writing. Which of these main points are emphasized in assessment depends on the instructional emphasis allocated to that area. Thus, writing literacy is a complex field of study. Similarly, valid assessment in this area demands an innovative and extensive undertaking if the results are to be worthwile.

Views on writing literacy. An intensive review of the literature has indicated that little research has been conducted in the area of the assessment of writing literacy. Bullock (1974) summarized the status of assessment of writing literacy in Great Britain by saying no national attempt had been made to monitor standards of achievement in writing. He stressed the importance and difficulties of validly measuring writing. The conclusion of the report is that writing can be adequately tested only by the scrutiny of a number of examples in which the child has had to cope with a variety of demands. The report further indicated that an important measure of success in writing was to differentiate between the styles appropriate for particular purposes. Writing literacy assessment should include writing tasks that are autobiographical and narrative as well as tasks that are explanatory and descriptive.

One notable exception to this lack of investigation of written literacy assessment was the national study conducted by National Assessment of Educational Progress (1970, 1972, 1975). Their seven year project is indicative of the recent importance placed on the integrated assessment of writing skills and activities in the United States. Slotnick (1973), former assistant director of NAEP, reviewed some of the implications for teaching of writing which have resulted from NAEP's assessment in writing. A major conclusion was that the best writing is produced when students are asked to write about subjects they find interesting or important and that the form of the task has very little to do with the quality of the response. Slotnick further argued that the importance of writing to the writer and his communication needs should be taken into account before deciding what to teach him. Maxwell (1973) replied to this discussion and concluded that since only rough drafts were assessed by NAEP, true writing assessment had not actually occurred. He suggested that revision of original drafts should be made an integral part of the total writing process in the teaching of writing.

Current viewpoints in the area of writing literacy assessment have established certain general criteria for valid assessment. On the basis of frequency of mention in the literature, Foley (1971) listed them as coherence and logic, development of ideas, diction, emphasis, and organization through sentence structure and paragraphing. Assessment of writing literacy then, to be valid, must attend to these or similar components. The studies discussed in the next section attempt to meet these types of criteria in several different ways. With more sophisticated use of computers in grading, increasingly comprehensive assessment of written literacy will be undertaken.

Assessment of writing literacy. The basic emphasis in the assessment of writing in the past focused on the problems of poor grammar, irregular spelling, and non-existent punctuation. Deductions for these problems from the score awarded for the positive features of a piece of writing often resulted in meaningless total scores that could not be interpreted. Awarding marks for good syntactical structure and deducting for errors has tended to bypass one other major component of writing. Was the writing successful in accomplishing its purpose? Had the writer been successful in persuading, clarifying, creating a mood, or developing and resolving a conflict? That is, had the writer successfully communicated through his writing? Also, subjective marking of essays in the traditional manner created numerous measurement problems. Often the validity and reliability of the marking scheme used was not substantiated. Current work based on these beginnings has resulted in some interesting new procedures being used in the assessment of writing literacy. Present project procedures have done much to remove the traditional validation and reliability problems found in older studies.

An approach to the assessment of writing literacy which includes the measurement of "success in communication" has been developed by National Assessment of Educational Progress (NAEP) and can be adapted for application in any school system. Essentially, this plan used five types of analyses of writing compositions. Because all grading was done by readers specifically trained in the scoring techniques involved, inter-judge reliability was high and comparability of grading was more assured than would be expected in traditional approaches. The following points briefly outline t five types of analyses which were conducted as part of the study.

1. Papers were first read and scored for overall quality by two readers independently each judge assigning a score ranging from 1 to 8. Thus, assigned scores reflected quality relative to other papers rather than quality as determined by set criteria.

2. The second analysis was a mechanical error count where errors in specifically defined areas were tabulated for a given quality level paper. Areas that have been included were punctuation, spelling, capitalization, agreement, run-on sentences, awkward constructions, faulty paragraphs, and word choice. The average score of two independent readers was the error count for the paper. Thus, error scores were characteristic of papers at a given quality level but the papers did not necessarily receive an overall grade simply because they did or did not contain certain errors.

3. The third analysis was a characterization analysis or description of the essay's syntactical charactistics. The essays were characterized according to the nine

areas reviewed in the error count. Correct as well as incorrect aspects of the paper were indicated. The characterizations were intended to provide insights which made the error counts more meaningful.

4. The fourth analysis utilized a special computer program developed by Slotnick (1972, 1973). Six qualities of writing were analyzed: fluency, spelling, diction, sentence structure, punctuation, and paragraph development. The information resulting from this computer analysis provided additional background for interpreting the error counts and characterizations. Eventually this program could eliminate the need for a teacher-tabulated error count.

5. The fifth analysis was called the Primary Trait System for Scoring Writing Tasks. This analysis was conducted independently of the other four. The primary trait score 1, 2, 3, or 4 essentially indicated whether or not a sample of writing contained the traits it must have in order to accomplish its purpose. A scoring guide for the assessment of the plotted narrative type of writing was developed by a local Edmonton Public School Board teacher and is contained in the Writing Appendix as an example of the application of this procedure to a system writing assessment at the junior high level.

Utilizing these analyses, each paper received the following component scores rather than one composite or total score: the holistic or overall impression score; a count of clearly specified errors; a written description of the structure of the paper; and a score indicating the degree of success of the paper as written communication. This type of component analysis of essays allowed for (1) the viewing of a paper from three separate and varied approaches--overall face value of the paper, its technical accuracy, and its ability to communicate successfully; (2) the tabulation of a very detailed profile of the status of diction, sentence structure, punctuation and paragraph development; and (3) valid comparisons to be made between essays written in different years and therefore also the identification of writing trends over time. Exercises released from both the 1970-71 and 1973-74 assessments are now available. All 1973-74 exercises are accompanied by a documentation page and individual primary trait scoring guides. A "Writing Mechanics and Grammar Scoring Guide" is also included.

Overview. The assessment of writing has not been the main focus of literacy assessment, likely because of the scoring costs. However, current concerns about student writing abilities are causing much more attention to be given to the assessment of written literacy than in the past. The review of projects currently being undertaken in the area indicate that future assessments of writing literacy will be much more comprehensive than in the past. Assessment

generalizations that have been arrived at with respect to writing literacy are discussed below.

Current assessment approaches to writing literacy stress the independent assessment of different writing components rather than the composite assessment of a piece of writing. New assessment procedures are very concerned with increasing the validity and reliability of writing literacy assessment. The increased availability of mechanical aids to assist in the assessment makes this type of scoring more feasible than formerly for large scale studies.

At present, only exercises from National Assessment of Educational Progress exist as partially standardized materials to be used this way in the assessment of writing literacy. Some attempts are being made in Great Britain to develop similar materials. In the opinions of experts in the area, the available standardized commercial component tests do not provide a measure of the essential qualities of writing literacy.

Among the three areas of literacy, writing literacy has had the least work. Assessment of writing literacy in the past has been incomplete either because not all components were tested or results were reported as vague composite scores. However, current work such as is being completed by NAEP introduces an essential set of assessment tools into the area of writing literacy. Further work in this area should produce many exciting developments.

Functional literacy assessment

Definition. Arriving at an operational definition of functional literacy is difficult because it is not a solitary trait. Being functionally literate can be defined as having the basic reading and writing skills necessary to participate in a mechanized, automated society. Without these minimal skills an individual cannot successfully function in an independent mode.

Nafziger et al (1976), quoting the United States Office of Education (USOE), defined a functionally literate person as

> one who has acquired the essential knowledge and skills in reading, writing and computation required for effective functioning in society, and whose attainment in such skills makes it possible for him to develop new aptitudes and to participate actively in the life of his times.

A similar definition was given by Hillerich (1976) in his paper. Many other researchers in the area have utilized definitions which resemble this basic operational definition of functional literacy.

Views on functional literacy. The study of functional literacy is relatively recent. Many of the tasks assessed in this area seem novel and would not be considered part of traditional formal assessment. Nafziger et al (1976) classified available functional literacy tests as informal tests, standardized tests, or criterion-referenced tests. Although some non-standardized informal tests of functional literacy are available, they have not been included for discussion. While they were suitable for individual classroom testing, the non-standardized informal tests lacked generalizability to large scale testing situations. Another group of tests not included for discussion were the standardized adult reading tests which measured adult vocabulary, graded reading comprehension, spelling or word meaning. Such tests provided a means of measuring approximate adult academic reading levels using adult vocabulary but did not really measure basic functional literacy. Since neither the informal nor the standardized reading tests seemed appropriate for large scale assessment of functional literacy, discussions of functional literacy tests were limited to the criterion-referenced test category. Usually, in a criterion-referenced assessment of functional literacy, reading criteria were selected from common daily reading and writing activities. Reading and writing tasks related to each criterion were developed. Actual application forms, labels, traffic signs, recipes, legal forms, advertisements and other real-life tasks were used in order to simulate daily reading and writing activities.

Much of the assessment research of functional literacy has concentrated on adult samples. Samples of studies of this type have been conducted by Harris (1970, 1971), Northcutt (1975), Lichtman (1973), and Sticht (1972). Several researchers, however, have used adolescent samples in their studies. Some of these include National Assessment of Educational Progress Mini-Assessment (1975), McDonald (1974), Hansen and Hesse (1974), and Murphy (1975). The Hansen and Hesse study also deviated from the usual assessment of functional literacy studies in that Bormuth's cloze technique was used in the development of the test.

Adult illiteracy has recently been designated a major target area of the Right to Read program in the United States. The extent of their commitment to reduce adult illiteracy is reflected in a national goal of the Right to Read program: to eliminate functional illiteracy by 1980 among 90% of the population over 16 years of age. The desire to determine the effectiveness of programs designed to reduce illiteracy has led in turn to a need for instruments that measure functional literacy. Thomas (1976) has surveyed cross-Canada activities in literacy under World Literacy of Canada auspices. In this Canadian study, 37% of the out-of-school adult population over 15 years of age were found to have had no more than elementary schooling. Seven per cent, or just under one million persons, had had no more than grade four schooling. Thomas indicated that numerous Adult Basic Education programs and literacy activities have come into

operation over the past seven years in an attempt to remedy this situation. For the most part, these were Manpower-sponsored programs, school-operated extension courses, volunteer programs, or upgrading programs sponsored by various government agencies.

The educational emphasis on functional literacy has increased. In the future, more assessment efforts will be directed towards valid functional literacy assessment.

Assessment of functional literacy. The world problem of functional literacy has been discussed by many writers. Estimates regarding the extent of illiteracy have been made and different criteria have been used to define functional literacy. Articles which discussed the international situation include those by various writers in Australia, Maher (1972) and Cullen and Pickering (1971); Britain, Allen (1975), as well as the National Foundation for Educational Research in England and Wales (NFER); and the United States, Grant (1971), Caughran and Lindolf (1972), and the National Assessment of Educational Progress (NAEP). These articles have shown the widespread stress currently placed on functional literacy.

Functional literacy can be considered from two viewpoints. First, it can represent only the most basic reading and writing skills needed to participate in society at a survival level. Second, it can represent an individual's status on essential real-life skills. At this second level, the assessment tasks are more complicated than at the first level, but the emphasis is the same.

In a paper presented to the 1976 Annual Meeting of the American Educational Research Association (AERA), Ayer (1976) discussed some of the major considerations involved in developing a test of functional literacy. Areas to be considered were purpose of the test, implementation of the purpose, developing the items, field testing the items, item analysis of tested items and setting a criterion level for functional literacy. The actual content of each area would depend on the viewpoint from which that testing project was approaching functional literacy.

The projects discussed below are interesting examples of work completed in basic functional literacy. Harris et al (1970), under the direction of the National Reading Council in New York, developed a functional literacy test for use in a Survival Literacy Study. In essence, this test consisted of numerous application forms which respondents were asked to complete. Harris reported that, using a criterion of 90 per cent correct, 13 per cent of their sample was from marginally literate to functionally illiterate in terms of ability to perform these common literacy tasks. Caughran and Lindolf (1972) raised a number of questions regarding this study. Included was the major question as to whether literacy should be measured only on a basis of ability to complete application forms. A second survey conducted by Harris et al (1971) explored respondent's ability to

answer straight forward questions about newspaper employment advertisements. Ninety-two per cent of the total sample answered all questions correctly. Sticht et al (1972) designed Project Realistic to determine functional literacy levels for selected military jobs having civilian counterparts. Jackson (1972) presented a paper at the American Educational Research Association (AERA) titled the Development of a Collection of Adult Tasks for Assessment of Literacy. A second analysis of tasks required for minimal competency was prepared by the Maryland Department of Education (1972). This analysis suggested that a student leaving school should be able to follow directions, use references, properly fill in forms, and gain information. Northcutt (1974) identified six general knowledge areas in which adults must achieve minimal competency in order to be successful in our society. These areas were knowledge of the world of work, consumer economics, health, community resources, government and law, and transportation. Northcutt (1975), for the Adult Performance Level (APL) Project, after extensive research identified the following additional four skill areas: communication skills, computational skills, problem solving, and interpersonal relations. Lichtman (1973) described the development and validation of R/EAL (Reading/Everyday Activities in Life), a new test designed to overcome problems in assessing functional literacy among adults and adolescents. Various application forms, signs, recipes, newspaper articles, advertisements, and maps were used in this test. Instructions were by audio-tape. This test was rated by Nafziger (1976) as being good for examinee appropriateness, but only fair for measurement validity, technical excellence, and administrative usability.

A major American national study by Educational Testing Services conducted by Murphy (1973a, b, and c, 1975) was undertaken to (1) identify the reading tasks actually performed by adults in a regular day; (2) construct a set of reading tasks intended to sample the universe of reading tasks ordinarily encountered by typical adults in our culture; and (3) survey actual adult performance on this set of reading tasks. In this particular study, no attempt was made to define functional literacy. However, in identifying reading tasks performed by adults in carrying out their ordinary activities, it was assumed that such tasks would have a favourable return to the individual and to society in general. This study also investigated the relationship of functional reading tasks to decoding skills. It was clearly indicated that the ability of decode was a predictor of success on the sample of adult reading tasks. An interesting part of the study examined the relationship of the reading tasks to a set of cloze tests. Within a grade level, scores on the two tasks were moderately correlated. That is, persons who could complete certain cloze passages could probably also pass certain functional reading tasks. However, at each grade level, respondents who failed the 35% criterion on cloze tests did manage to pass a large number of adult functional reading tasks. The main purpose of this study was successful, that is, a set of functional tasks was

identified, tests were constructed to measure these tasks, and a survey of performance levels on these tasks was completed. However, the part of the study dealing with the cloze techniques was only partially successful, since the passages selected for the cloze techniques were found to be too difficult for assessing minimal functional reading competency. Further work would need to be done to select appropriate cloze passages.

Nafziger et al (1976) have recently published a listing and evaluation of currently available instruments used in the assessment of functional adult literacy. Nafziger et al discussed (1) some of the problems in defining and measuring functional literacy, and problems in choosing tasks to measure literacy; (2) activities undertaken to identify tests; (3) forty-one evaluative criteria used in evaluating the tests and the categorization of these criteria under measurement validity, examinee appropriateness, technical excellence and administrative usability; (4) descriptive reviews of the tests as grouped under criterion-referenced tests, standardized tests, and informal tests; (5) evaluations of these tests; and (6) an overall summary indicating some general strengths and weaknesses of different types of tests. This report indicated that adult literacy testing was still a developing field marked by instruments of varying quality.

Several researchers have conducted major investigations into the assessment of functional literacy using adolescent populations. Project staff at ETS have developed an experimental test for assessing reading competency in schools. In general, the reactions to this school-oriented test were favourable on the part of both teachers and students. McDonald and Mooreman (1974) developed assessment materials to measure individuals' status on real-life skills. Their criterion-referenced reading test, Minimal Reading Proficiency Assessment (MRPA), when administered on a district-wide basis, appeared to demonstrate high school graduates' ability to read at least at the ninth grade level of proficiency. Hansen and Hesse (1974) designed a study for similar use in a Wisconsin school district. In this study, however, the Bormuth cloze technique was directly applied to the development of a functional literacy test. This application provided a novel procedure in the measurement of functional literacy since most other measures used reading tasks directly found in everyday experiences. Hansen and Hesse found that between 25-30 per cent of the students leaving high school were able to gain little information from basic materials society expects them to be able to read.

National Assessment of Educational Progress (1975) conducted a national Mini-Assessment of Functional Literacy (MAFL) under the Right to Read program, designed to determine the extent of functional illiteracy among United States high school graduates. MAFL exercises were selected to present formats of reading materials frequently encountered in everyday life and were administered to a national sample of 17-year-olds in school in 1974. The following five formats

were used for items: passages; drawings, pictures, coupons; charts, maps, graphs; forms; and reference materials. Scoring analysis and data processing were done in conjunction with the NAEP regular Writing Assessment in 1973-74. The study generally indicated that 17-year-olds in the United States had gained in basic reading skills over the past five years.

Overview. The multifaceted nature of functional literacy has often been lost through the use of composite scores such as grade levels to indicate literacy or non-literacy. There appears to be a general consensus among the researchers who have investigated assessment of functional literacy since 1970 that any index used to measure literacy must be more than simply a grade level passed in school. Such an index must be based instead on an actual performance score derived from tests which measure common daily reading and writing activities. Some of the criterion-referenced tests employ an open-ended format, others the multiple-choice format, and others one of several cloze techniques. Those applying tests for actual determination of literacy levels seem to prefer the recently developed criterion-referenced tests such as the Functional Literacy Test, R/EAL, MRPA, MAFL, or the tests developed by ETS in their Adult Functional Reading Study. All of these employ forms, signs, letters, recipes, instructions, and other daily encountered tasks. Current public interest into the practical aspects of functional literacy assessment will likely promote increased use of this type of testing to determine if students completing school and adults functioning in society have been equipped with the necessary survival skills in literacy.

SUMMARY OF RESEARCH FOR ORAL COMMUNICATION AND LISTENING ASSESSMENT

The language arts program presented in its most progressive stance involves more than reading and writing literacy. Within the area of verbal skills, speaking and listening must interact with reading and writing to produce the necessary experiences, thoughts and feelings that develop the required vocabulary and syntax for clear, effective communication. Both of these skills have been present in instruction for many years but only recently have they started to share part of the major focus of the language arts program. Therefore, the receiving and expressing components are discussed further in the next two sections.

Oral communication assessment

Definition. Oral communication can be defined as a spoken unique communication that conveys an idea, feeling or experience to others. The speaker must have in mind the effect to be achieved, the nature of the audience he is attempting to communicate with, the appropriate form of communication to use and the material to be communicated.

According to Kennedy and Moore (1971), in elementary school the primary instructional emphasis with respect to oral communication should be upon basis speech skills in functional situations, not upon formal public speaking. In junior and senior high school, emphasis shifts to making more formal presentations such as speeches and debates. A valid assessment program would necessarily reflect the different emphasis placed on oral communication at various age levels.

In more technical terms, oral communication can be divided into two areas, basic speaking skills and speaking activities. Basic speaking skills include attention to articulation, enunciation, pronunciation and voice. Speaking activities include conversational speech, reporting, story telling, dramatizing, and so on. Comprehensive assessment of oral communication would need to attend to both of these areas. The present state of assessment of oral communication has tended to look at individual aspects of either basic speaking skills or speaking activities, rather than attending to a comprehensive whole. The next section will look at some differing opinions on assessment of oral communication.

Views on oral communication. Oral communication has been considered with different degrees of importance in the formal educational structure. At one extreme, in 1755, Samuel Johnson wrote that spoken language was not worth studying. Today, oral communication seems very desirable as a subject for study. The actual content of oral communication and how it should be assessed is still not completely clear. Wilkinson (1969) emphasized this point by

questioning the absence of a comprehensive theoretical framework for oral communication and its subsequent assessment. His work attempted to define concepts in oral performance, linguistic studies and practical oral presentations, the measurement of which would lead to valid and effective assessment of oral communication.

Studies of oral language, particularly spontaneous utterances, have been hampered by the complexity of analysis procedures required. Graham (1971) pointed out that even with modern electronic equipment which makes the collection of language output relatively easy, there still exists the practical and theoretical problems of how to segment, analyze and quantify the significant properties of oral language samples. A second concern of any oral language measurement is the degree of subjectivity involved in the assessment. The Bullock Report (1975) cautioned against using extremely artificial oral language assessment situations as the results may not truly reflect the student's ability to converse in a natural situation. Questions that the Bullock Report feel valid oral language assessment would answer were as follows:.

1. Is talk being promoted as an instrument for learning and thinking?
2. Has the pupil's command been extended over the varied resources of spoken language?
3. Does the pupil look upon discussion as an activity in which it is as important to interpret and accommodate the views of others as to express his own?
4. Is the pupil demonstrating the capacity to move from the concrete to the abstract, and from the immediate to the distant and hypothetical?

A third concern involves the time availability in our present teaching structure, and the lack of qualified classroom personnel to undertake extensive teaching or assessment of oral communication. Recent developments in language arts programs are attempting to remedy this deficiency by promoting the teaching of oral communication skills and activities within the classroom. The current Alberta Provincial Curriculum (1973) stresses not only written communication, but also oral communication as being essential in a child's language arts development.

Assessment of oral communication. Traditionally, oral communication was not assessed as a separate component of an English program, particularly with respect to basic oral skills. Current research does indicate, however, that work is being attempted in both the assessment of basic oral skills and oral activities. Initially the interest originated with the improvement of second language assessment materials, but has recently been applied to regular language arts programs as well. In many instances, modifications to second language assessment materials have made them

appropriate for use with regular student populations in language arts. In this section, findings relevant to the assessment of basic oral skills and oral activities are included.

A comprehensive piece of work involving basic oral skills was reported by Carrow (1974). Carrow's test, Elicited Language Inventory, consisted of sentences of various lengths and degrees of complexity. After hearing the sentences, children were asked to imitate them and their responses were recorded on tape. The scoring of the sentences indicated the developmental status of a child's base sentence construction and grammar. Since the sentences covered a wide range of complexity, it was possible to examine what the child was saying, as well as to make comparisons with adult speech. One main feature of this instrument, in the author's opinion, was that the sentences paralleled the child's spontaneous speech patterns. A similar testing procedure was developed for use with younger children by Graham (1971). He used a test consisting of sets of pictures, each picture set designed to elicit a particular grammatical transformation. Maximum cueing was used to encourage the subject, if it was within his competence to do so, to produce particular verbal structures which might not have occurred in spontaneous utterances in normal surroundings. According to Graham, a possible future use of this testing technique is the investigation of patterns of language deficiencies in primary school children. However, since both Carrow's and Graham's test instruments require extensive individual administrations, they are of limited value to the classroom teacher.

Following are two descriptions of assessment techniques that would be more manageable in a regular classroom setting for the assessment of basic oral skills. Samples of this type of testing materials were illustrated by Cayden (1975) and Dobbyn (1976). Cayden was concerned with possible distortions caused by using a standardized testing instrument rather than a more natural setting when testing basic oral skills with a group of primary children. His work revealed that a standardized testing instrument caused students to make more errors than they normally made in a natural conversation situation. A possible explanation was that the more formal testing situation produced an unnatural set in speech patterns when children were faced with many examples of a particular speech form at one time. Cayden stressed the need for natural oral testing situations for young children. Dobbyn, in another classroom investigation with older children developed an objective test of pronounciation that was quick to administer and gave reliable results in a regular classroom situation. An interesting aspect of his work was that his procedures could be applied to the assessment of other basic oral skill areas. Using this method, teachers would be able to collect required information about oral communication skills and thus structure their oral language arts programs from their findings.

More research has been carried out in the oral activities section. One typical type of assessment of a student's ability to describe or give directions in a structured situation was described by Williams et al (1970). They asked students sets of questions based on three selected pictures. The oral reactions were taped and analyzed to determine differences in oral language levels of elementary school children from grade 1 to grade 6. A similar study was structured by Rose et al (1973) to assess the quality of inter-communication among young children. The task that each receiving child was required to undertake involved the placement of objects onto a pegboard from verbal instructions such that it would match a preset pegboard visible only to the presenting child. This study showed that with an increase in age, both the communication proficiency and linguistic proficiency improved. This technique could be used both clinically and in a classroom to identify characteristics of competencies that lead to adequate and effective communication, as well as to diagnose similar competencies that are required as part of a language arts program. A slightly different approach was used by Melear (1973). Melear felt that freedom of expression was essential in any attempt to estimate the quantity or quality of a child's real language ability. He adapted work in the oral language area to construct a measurement procedure titled the Informal Language Inventory. This procedure involved having a group of children draw a picture and then individually volunteer to describe their picture. The description was recorded and analyzed for grammatical content. The administration of the instrument was successful in locating deficiencies in vocabulary and grammatical structure. Simple modifications to this technique could easily provide a more comprehensive tool for the assessment of oral language.

Other investigators have been concerned with the effectiveness of students' oral presentation. Robinson (1971) attacked the problem by attempting to determine the advantages and disadvantages of differing amounts and kinds of student participation in oral communication testing tasks. His work stressed the use of at least some student participation for valid testing to take place. Sample testing exercises for different degrees of participation were provided in his paper. Folland and Robertson (1976) constructed a testing situation very similar to that advocated by Robinson. Their concern was to assess objectively oral conversation. The procedure involved a limited, properly oriented group of students who participated in free discussion initiated by a pre-recorded tape. Two English teachers did not participate in the discussion, but scored the students as they contributed to the discussion. The authors determined that this approach was much more reliable for assessing conversational skills than the traditional one-to-one interview approach, and that it could be usefully and easily adapted for use in other situations. Willinson (1969) also studied spoken language and ways of improving a child's presentation. Further work by Skull and Willinson (1969) resulted in a reliable and valid method to score oral presentation. Tucker (1970) presented an

overview of important factors to remember when assessing either single or group communications.

Work undertaken by Stauffer and Pikulski (1974) indicated that the more practice young children have in using oral skills, the more proficient they become. Child-dictated stories from each child were recorded throughout the year. By the end of the year, their stories included both basic sentences and longer sentences reflecting the use of modifiers, clauses, and other elaborations. Student stories dealt with a wide variety of topics, and a total of 3,331 different words were tabulated. Research on a more formal type of presentation at an older student age was undertaken by Potter and Strother (1975). These authors developed and reported on an accountable procedure for measuring and recording counter-productive verbal and nonverbal responses which were presented by a student in an oral production. This procedure shows promise for objective continuous measurement of both student progress in oral language and the effectiveness of the speech development program.

Overview. Until recently oral communication was not assessed as a separate component in the language arts program. As a result of current emphasis in the oral language arts area, assessment investigations have been undertaken in both the areas of basic oral skills and oral activities. Several generalizations arise from the review of studies of both these areas. Investigators emphasize the importance of controlled but natural testing situations. It has been shown that artificial testing situations may distort test results, particularly with younger children. The lack of standardized commercial instruments for measuring oral language competencies is evidence of the special problems associated with this type of assessment.

To this point, the most successful testing has taken place on an individual or small group basis. Further modifications to some of the suggested procedures would make it possible to use these assessment techniques with one test administrator in a classroom of normal size without destroying the test's validity. It is assumed in the above generalizations that the appropriate audio equipment would be available.

It appears from the review of the literature that investigators are striving to measure not the mechanical manipulation of form or content, but rather originality in form and content. In other words, the research emphasis is aimed at higher cognitive levels rather than recall activities. Much further work remains to be done in the area to bring it into a practical form for generalized use in the schools.

Listening assessment

Definition. Listening can be defined as the act of receiving oral language. Individuals must listen to find out how,

where, and what to buy; what is happening in the world; what is available for entertainment; and so on. In today's world, the spoken word is the most powerful medium of communication and therefore effective listening skills are essential.

Responsible listening demands more than attention to the extrinsic conditions which make auditing possible. After the stimulus has been received, it becomes the active function of the listener to interpret, evaluate and respond to the message, meaning and mood of the sender. Students must learn the necessary skills to carry out this type of responsible listening. A valid assessment program would necessarily reflect these different components of a listening program.

In more technical terms, listening can be divided into two areas: basic listening skills and listening activities. Basic listening skills include such components as auditory acuity and auditory discrimination. Listening activities include listening to stories and poetry, to reports, lectures and talks, and to conversations and discussions. Comprehensive assessment of listening would have to attend to both of these areas. The present state of assessment for listening has tended to look at individual aspects of either basic listening skills or listening activities, rather than attending to a comprehensive whole. The next section will look at some differing opinions on listening.

Views on listening. Educators within the field of reading have long emphasized the importance of listening skills with respect to reading and have contended that a measure of listening comprehension should be predictive of reading achievement. As a result of this, in recent years there has been an increasing emphasis on the place of listening in education. Wilkinson (1970) stated that curriculum construction and assessment should be thought of both in terms of written literacy and "oracy" (oral literacy) which includes both speaking and listening. His work in the area has attempted to promote this viewpoint.

Piaget (1963) had put forth similar views. His work demonstrated the importance in a child's development of progressing from language sounds to internal articulation with respect to the sounds. This to Piaget was one of the first developmental factors of true intelligence in the child. The effective use of listening, not just hearing, is necessary for the child to gain the fullest meaning of what is present in his auditory field. Instruction can be concerned with listening in that it can help the individual learn to structure his hearing in order to attend to those stimuli considered most important in obtaining the proper meaning from a given situation.

Despite the importance of this skill, research suggested that individuals listen very badly. For example, studies summarized by Wilkinson and Stratta (1972) showed that students listening to

lectures have been found to comprehend half or less than half of the lecture material. Possible reasons given by Wilkinson and Stratta were the lack of interest present in the lectures and the passive role required of the listener. The existence of internal barriers within the receiver to information being received from a sender was listed by Williams (1974) as an additional reason for poor listening.

Although considerable work has been undertaken to determine the relationship between reading and listening, little has been done on speaking and listening. The degree of significant relationship between speaking and listening is in dispute at the present time, with studies giving conflicting results. Wilkinson (1970) summarized these results and indicated the need for further research in the area.

Current viewpoints in the area of listening assessment have established certain general criteria for valid testing. The more important among these are summarized as follows:

1. Listening tests should be representative of the listening skills the individual needs for the varied activities in which he is daily involved. It has been found that measurements from artificial testing situations bear little resemblance to actual listening behavior (Bullock, 1975).
2. It is not advisable to ask subjects to listen to just one piece of English and judge their listening ability from their understanding of that alone, since the subject matter may not appeal to them, or they may be distracted at a crucial moment which would influence their understanding (Wilkinson et al, 1974).
3. In choosing test material, it seems important that this material should be presented to students in a context, not a set of isolated items (Wilkinson et al, 1974).
4. It appears crucial that stimulus selections for listening tests should be presented from standardized tapes to avoid examinees being given an advantage from examiners' cues, or a disadvantage from examiners' improper articulation (Wilkinson et al, 1974).

Any measurement of listening must attend to these guidelines for quality listening assessment to exist. Any measurement, to be valid, should consider as many facets of listening as possible. In an interesting hierarchial arrangement of listening skills, Weaver and Rutherford (1974) have attempted to consider as many of these skills as possible at the elementary level. They have divided listening skills into environmental, discrimination, and comprehension skills. Any assessment program in listening must take into account the influence of environmental skills, and at the same time provide measurement of both discrimination and comprehension skills.

Assessment of listening. Prior to 1950, listening was not assessed as a separate component of an English program. One of the first listening comprehension tests was developed in 1953 by Brown and Carlson. They constructed a taxonomy of listening skills where they classified receptive and reflexive skills. Subsequent investigators tended to build on the original work of these two authors until the early 1960's. According to Wilkinson and Stratta (1970), what these investigations lacked was an interest in language and the essential accompaniment of it, an interest in situation. The first lack was displayed in the literary nature of much of the language in their tests--it was written language read aloud. The second lack was discerned in the fragmentary nature of many of the test questions. In a real situation there always is a context to make clear which are the important points. Thus, current assessment in listening should be concerned with genuine spoken English as much as possible. In this section, findings relevant to the assessment of basic listening skills and listening activities are discussed. In many of these, the above criteria for quality listening tests have been taken into account.

One question of interest to researchers has been to determine whether or not listening skills can be taught. A study by Penfield and Marascuilo (1972) was structured to answer this question. They wished to determine whether or not it was possible to train students to become better listeners relative to a specific set of skills developed by Hackett (1970). Results showed that training was effective at the grade 2 level, but as students became older (grade 5) environmental conditions played a stronger and stronger role in the acquisition of information via listening. A similar question was studied by Rosner (1974) with respect to much younger children. With Rosner's group of preschoolers, the data clearly indicated that auditory analysis skills could be taught. The results supported the argument that it is feasible to teach auditory analysis skills well in advance of reading instruction, thus dealing effectively with at least one aptitude that is closely related to reading achievement. Palnatier and McNinch (1972) further substantiated the results of the Penfield and Marascuilo study. Palnatier and McNinch found that no significant growth in listening ability occured among high school students who were exposed to a study skills program in which listening growth was an expected outcome.

A new dimension in the assessment of speech sound discrimination was developed by Goldman et al (1971). After reviewing available tests for preschool and primary children, the authors decided there was need for a test to evaluate speech sound discrimination ability. The test which they developed minimized the influence on students' scores of variables such as abstractness of materials or familiarity with vocabulary and illustrations used. The scoring of this individually administered test was intended to identify individuals with deficiencies in auditory attention. Further validation of the Goldman-Fristoe-Woodcock test was undertaken

by Finkenbinder (1973). His findings questioned the value of the test as a predictive instrument of reading ability in the primary grades. Several reviewers in Buros have discussed the advantages and disadvantages of this test. Although the general consensus was that the test had merit, questions remain concerning some aspects of its validity and reliability, indicating further research on formal testing instruments of this type is necessary.

Three studies which investigated more theoretical aspects of listening skills were those completed by Mecham (1971), Rosner and Simon (1971) and Schwartz and Goldman (1974). Mecham described the development of a testing technique to measure verbal listening accuracy in children in either a group or individual situation. The test which was demonstrated to be reliable, required students to identify a picture for a given verbal label. In their study, Rosner and Simon developed an individually administered auditory perception test for elementary children. Results indicated that listening improved in step with maturation and improved reading skills. Schwartz and Goldman investigated the effect of variations in the context of stimulus-item presentation on speech-sound discrimination under different background listening conditions. Their work showed that at each age level significantly more errors were made in the paired comparison context than in carrier phrase or sentence context. This suggested that some aspect of this manner of assessing speech-sound discrimination adversely influenced performance. Other related findings were also reported. In summary, these studies indicate again the need for further investigations to prepare valid, comprehensive and formal assessment instruments of listening skills.

Other researchers have concentrated on the listening activities area emphasizing the measurement of listening comprehension skills. Butler (1970) examined the listening of sixth grade children. Using the STEP listening test as h[illegible] assessment instrument, he found that good listeners were found to ha[illegible] significantly higher scores on standardized tests of word meaning, paragraph meaning, and intelligence than poor listeners. Working with younger children, Nesbitt (1968) and Wallner (1974) investigated listening characteristics of first grade pupils. Nesbitt showed that a majority of first grade pupils through listening were able to understand a selection designed for use at the second grade level, and that almost a third of these students were able to understand a selection designed for the fifth grade level. This study indicated the value of developing listening skills in younger students since it appears that through listening they were able to comprehend materials well above their own reading level. Wallner was concerned with developing a test of listening comprehension that would validly and reliably measure both literal and inferential comprehension among first grade students. Two forms were developed which were shown to have sufficient predictive validity with reading achievement to warrant their use for educational purposes.

At the grade 3 level, Kennedy and Weener (1973) conducted a study to determine the effectiveness of individualized training with the cloze procedure to improve reading and listening comprehension. Among their findings, it was shown that the cloze procedure could be used to measure listening comprehension. Bowdidge (1967) had shown previously that a cloze listening test could be used with high school students. The group administered test that he developed was constructed to measure recall of specific information, ability to grasp the thought of a passage as a whole and ability to apply various contextual clues. This test appears to have some value as a group administered instrument for assessing comprehension aspects of listening ability.

A more informal component of listening assessment was described by Williams (1972). He examined the use of Listening and Speaking Self-Evaluation Charts to help students determine what they hear and say. He suggested that use of self-evaluation charts would help students develop critical thinking skills. This technique emphasized assessment in a natural setting. The use of natural settings in any form of listening assessment has been the major thrust of work undertaken in Great Britain. In the more informal types of assessment appropriate for the classroom setting, it is not difficult to incorporate the criteria of natural setting as suggested by Williams. More problems are encountered, however, when one moves to a more formal type of assessment. Wilkinson et al (1974) dealt with these problems very effectively in the construction of three Listening Comprehension Assessment batteries for children aged 11 to 18. One of the major aims of the construction of the three test batteries was to avoid the use of written English read aloud, and to concentrate on using material that people are hearing naturally most of the time, that is, spoken language. Wherever possible, natural conversation that had been recorded on tape served as a basis for test passages. The authors stressed that the emphasis in their listening tests was on the richness and variety of language rather than upon skills. As such, they would see at least part of the test's purpose as the development of an awareness of the features and functions of spoken language.

Overview. The assessment of listening as a separate component is a recent innovation of language arts programs. As a result of current emphasis in the listening area, assessment investigations have been undertaken in both the areas of basic listening skills and listening activities. After reviewing studies in both areas, generalizations with respect to assessment arise.

One of the most important generalizations is the use of controlled but natural materials in the construction of listening tests--use of spoken English as opposed to written English. It has been shown that where written English is used on listening tests, test results do not correlate with testees' actual listening performance as observed in a non-test situation. Continuing this

generalization, it has been shown that test results are more valid when materials used are placed in context. This concept opposes some of the techniques in the early development of listening tests. Most research reported to date in the area of listening assessment makes use of and assumes the availability of appropriate audio equipment. Many researchers stress the use of taped presentations to ensure standardization from administration to administration.

Few commercial tests currently exist for the assessment of listening. Two major efforts have been the test battery developed by Goldman et al in 1970 which measures aspects of listening skills, and the three batteries developed by Wilkinson et al in 1974 which measure comprehensive aspects of listening activities. The most promise for formal assessment in listening lies in the further development of test batteries such as these.

As a language arts component, listening has had little notice. However, in the modern world, the ability to listen effectively is becoming an increasingly important and valid educational activity. Thus, the development of valid and reliable testing instruments to evaluate both student and program progress becomes essential.

BIBLIOGRAPHIES

INTRODUCTION TO BIBLIOGRAPHIES

As the three areas of communication skills - literacy, oral language and listening were researched, many books and articles from throughout the world were reviewed. Many were discarded as not relating directly to assessment of regular elementary and secondary students while others were felt worth listing in the bibliography as possible additional sources but were not written up in the appendices. Summaries of research and the most important studies were written up in the summaries of research and the introductory reports. Fewer than five percent of the actual studies listed are not available from the Pupil Assessment Branch, Edmonton Public School Board and these are available from the University of Alberta, Edmonton, Alberta.

Listings are grouped in this bibliography in six sections. The first section lists general material used in the introductory report. Literacy is divided into three sections - reading, writing, and functional. This is followed by a listing for oral language and for listening.

GENERAL BIBLIOGRAPHY

BUROS, Oscar K. (ed.), Reading Tests and Reviews. The Gryphon Press, Highland Park, New Jersey, 1968.

BUROS, Oscar K. (ed.), The Seventh Mental Measurements Yearbook, Vol. I & II. The Gryphon Press, Highland Park, New Jersey, 1972.

Elementary Language Arts Handbook, Interim Edition, Government of Alberta, Department of Education, 1973.

These publications were referred to in "Assessing Communication Skills: An Overview" but not directly in the five summaries of research. Other authors referred to in "Assessing Communication Skills: An Overview" whose work relates directly to one of the five communications are listed in the appropriate bibliographies following.

BIBLIOGRAPHY FOR READING ASSESSMENT

AARON, ROBERT L., Using the Cloze and Sight Vocabulary to Develop A Group Informal Reading Inventory. New York, New York: Twentieth Annual Meeting of the International Reading Association, 1975. 12pp. (Unpublished Paper). ED109627.

AMBERDEN, N., "Don't Leave Reading to the Psychologists," The Teacher, 26, No.26, June, 1974.

ALVORD, DAVID J. and BARBARA BRITTINGHAM, "Evaluating Performance on National Assessment Objectives: Norm-Referenced and Criterion-Referenced Interpretations," The Journal of Educational Research, 68, No.2:59-61, October, 1974.

ANDERSON, RICHARD C., "How to Construct Achievement Tests to Assess Comprehension," Review of Educational Research, 42, No.2: 145-170, Spring, 1972.

BADCOCK, MAX, "Literary or Literate," The Educational Magazine, 31, No.1:5-9, 1974.

BALYEAT, RALPH, and DOUGLAS NORMAN, "Lea-Cloze Comprehension Test," The Reading Teacher, 28, No.6:555-56[illegible], March, 1975.

BICKLEY, A. C., BILLIE J. ELLINGTON, and RACHEL T. BICKLEY, "The Cloze Procedure: A Conspectus," Journal of Reading Behavior, No.2:232-249, 1970.

BOOKBINDER, G. E., "Should We Believe What the Reading Surveys Tell Us?," Reading, 7, No.2:14-21, June, 1973.

BORMUTH, JOHN R., "Factor Validity of Cloze Tests as Measures of Reading Comprehension Ability," Reading Research Quarterly, IV, No.3:358-365, Spring, 1969.

BORMUTH, JOHN R. Development of Standards of Readability: Toward A Rational Criterion of Passage Performance. Chicago, Illinois: Chicago University, June, 1971. 219pp. (Final Report). ED054233.

BORMUTH, JOHN R., "Reading Literacy: Its Definition and Assessment," Reading Research Quarterly, 9, No.1:7-66, 1973-74.

BULLOCK, SIR ALAN. A Language for Life: Report of the Committee of Inquiry Appointed by the Secretary of State for Education and Science Under the Chairmanship of Sir Alan Bullock F. B. A., Her Majesty's Stationery Office, London, 1975.

BURKE, ELIZABETH, "Standards of Reading: A Critical Review of Some Recent Studies," Educational Research, 17, No.3:1-12, June, 1975.

CARVER, RONALD P. Rules for Developing Reading-Input and Reading-Storage Tests. Kansas, Mo.: University of Missouri, 1970. (Unpublished Manuscript).

CARVER, RONALD P., "A Computer Model of Reading and It's Implications for Measurement and Research," Reading Research Quarterly, 6:449-471, 1971a.

CARVER, RONALD P., and C. A. DARBY, "Development and Evaluation of a Test of Information-Storage During Reading," Journal of Educational Measurement, 8, No.1:33-44, 1971b.

CARVER, RONALD P., "Reading as Reasoning: Implications for Measurement," Assessment Problems in Reading. Newark, Del: International Reading Association, 1973.

CARVER, RONALD P. Improving Reading Comprehension: Measuring Readability. Washington, D. C.: American Institute for Research, May, 1974a. 90pp. (Final Report). ED092920.

CARVER, RONALD P., "Measuring the Primary Effect of Reading: Reading-Storage Technique, Understanding Judgments and Cloze," Journal of Reading Behavior, VI, No.3:249-274, 1974b.

CARVER, RONALD P., "Comparing the Reading-Storage Test to the Paraphrase Test as Measures of the Primary Effect of Prose Reading," Journal of Educational Psychology, 67, No.2:274-284, 1975a.

CARVER, RONALD P., "Revised Procedures for Developing Reading-Input, Materials and Reading-Storage Tests," Journal of Reading Behavior, VII, No.2:155-172, 1975b.

CARVER, RONALD P., "Further Research on the Reading-Storage Test as a Measure of Gain During Reading," Journal of Reading Behavior, VII, No.4:401-414, Winter, 1975c.

DOWNING, JOHN and E. C. DALRYMPLE-ALFORD, "A Methodological Critique of the 1973 IEA Survey of Reading Comprehension Education in 15 Countries," Reading Research Quarterly, 10, No.2:212-227, 1974-75.

FARR, ROGER. Reading Achievement in the United States: Then and Now. Educational Testing Service, Princeton, N. J., Aug., 1974. 174pp. ED109595.

FINN, PATRICK J., "A Question Writing Algorithm," Journal of Reading Behavior, VII, No.4:341-367, Winter, 1975.

FORBES, ROY H., "NAEP: One 'Tool' to Improve Instruction," Educational Leadership, 34, No.4:276-281, January, 1977.

GOODACRE, ELIZABETH J., "Reading Research in Britain-1973," Reading, 8, No.1:14-17, March, 1974.

GOODACRE, ELIZABETH J., "Reading Research in Britain-1974," Reading, 9, No.1:5-11, March, 1975.

HARRIS, ALBERT J., and MILTON D. JACOBSON, "Predicting Twelfth Graders' Comprehension Scores," Journal of Reading, 20, No.1:43-46, October, 1976.

HUSEN, TORSTEN,(ed.),International Association for the Evaluation of Educational Achievement. IEA Six-Subject Survey Instruments: Literature Tests. Stockholm, 1974. 55pp.

HUSEN, TORSTEN,(ed.),International Association for the Evaluation of Educational Achievement. IEA Six-Subject Survey Instruments: Reading Comprehension Tests. Stockholm, 1974. 72pp.

Inner London Education Authority. Literacy Survey: Summary of Interim Results of the Study of Pupil's Reading Standards. Proposals for Further Action. London, 1969. 21pp. (Mimeographed).

Inner London Education Authority. Literacy Survey: 1971 Follow-Up-Preliminary Report. London, December, 1972. 11pp.

Inner London Education Authority. Literacy Survey: 1973-74 Follow-Up-Preliminary Report. London, November, 1975. 11pp.

JAMES, P. C., "The Use of Cloze Procedure with the ESN Children," Reading, 9, No.1:26-33, March, 1975.

KENNEDY, DOLORES, "The Cloze Procedure Use It to Develop Comprehension Skills," Instructor, p.82-86, November, 1974.

KLARE, GEORGE R., "Assessing Readability," Reading Research Quarterly, VX, No.1:62-98, 1974-75.

LEDBETTER, J. T., and HARVEY DANIELS, "Forum: Is There a Decline in Literacy?," English Journal, 65, No.6:16-20, September, 1976.

LETHBORG, NOEL, LEIGH NAUNTON and JACKIE STANYON, "Literacy at Merrilands High School," The Educational Magazine, 31, No.1:22-25, 1974.

McCULLOCH, CONSTANCE M. et al, "Discussion Review of Reading Comprehension in Fifteen Countries," American Educational Research Journal, 11, No.4:409-414, Fall, 1974.

NEVILLE, MARY H. and A. K. PUGH, "Context in Reading and Listening: A Comparison of Children's Errors in Cloze Tests," British Journal of Ed. Psychology, 44, Part 3:224-232, November, 1974.

NEVILLE, MARY H., and A. K. PUGH, "Context in Reading and Listening: Variations in Approach to Cloze Tasks," Reading Research Quarterly, XII, No.1:13-31, 1976-77.

OLSON, MARGOT A., "Criterion-Referenced Reading Assessment in a Large City School District," Journal of Reading Behavior, VII, No.4:387-396, Winter, 1976.

O'REILLY, ROBERT P. An Objective Procedure for Constructing Domain-Referenced Tests of Literal Comprehension. New York State Education Department, October, 1975. 1pp. (Mimeographed).

OSGOOD, CHARLES E., "On Understanding and Creating Sentences," American Psychologist, 18:735-751, March, 1963.

PAGE, WILLIAM D., "The Post Oral Reading Cloze Test: New Link Between Oral Reading and Comprehension," Journal of Reading Behavior, VII, No.4:383-389, Winter, 1975.

PETERSON, JOE, E. PARADIS and N. PETERS, "Revalidation of the Cloze Procedure as a Measure of the Instructional Level for High School Students," a chapter in Twenty-Second Yearbook of the National Reading Conference, Vol.I, Phil L. Nache, ed., Boone, North Carolina: National Reading Conference Inc., 1973.

PYRCZAK, FRED, "Passage Dependence of Items Designed to Measure the Ability to Identify the Main Ideas of Paragraphs: Implications for Validity," Educational and Psychological Measurement, 34, No.2:343-348, Summer, 1974.

PYRCZAK, FRED, "Passage-Dependence of Reading Comprehension Questions: Examples," Journal of Reading, 18, No.4:308-311, January, 1975.

PYRCZAK, FRED. Context-Dependence of Items Designed to Measure Ability to Derive the Meanings of Words From Their Context. San Francisco, California: National Council on Measurement in Education, April, 1976. 11pp. (Mimeographed).

RAIG, LUCY, "Whose English Do You Speak?," The Educational Magazine, 31, No.1:14-17, 1974.

ROBINSON, RICHARD D. An Introduction to the Cloze Procedure: An Annotated Bibliography. Newark, Deleware: The International Reading Association, 1972. 12pp.

ROBINSON, RICHARD D., "The Cloze Procedure: A New Tool for Adult Education," Adult Education, 23, No.2:87-99, 1973.

SCHERICH, DR. HENRY, and DR. GERALD HANNA. Passage-Dependency Data in the Selection of Reading Comprehension Test Items. San Francisco, California: National Council on Measurement in Education, April, 1976. 10pp. (Mimeographed).

SCHWILLE, JOHN and SUSANNE MARSHALL. The IEA Six-Subject Data Bank: A General Introduction. Stockholm: International Association for the Evaluation of Educational Achievement, 1974. 66pp. (Mimeographed).

SHAFER, ROBERT E., "A National Assessment in English: A Double Edged Sword," Elementary English, 48:188-95, April, 1971.

SHERWOOD, PHILIP, "Evaluation, The English Experience," The National Elementary Principal, 56, No.1:39-42, September/October, 1976.

SPALDING, HEATHER, "Radical New Developments in Assessment," The Teacher, 26, No.26, June, 1974.

STOCKWELL, R. P. et al. Integration of Transformational Theories on English Syntax. Los Angeles, California: University of California, 1968. (Unpublished Paper).

THORNDIKE, ROBERT L. International Studies in Education III: Reading Comprehension in Fifteen Countries An Emperical Study. Stockholm: Almquist and Wiksell, 1973. 179pp.

TUINMAN, J. Inspection of Reading Comprehension Passages as a Function of Passage Dependency of Test Items. Bloomington, Indiana: Indiana University. Institute for Child Study, 1972. (Research Report).

TUINMAN, J., "Determining the Passage Dependency of Comprehension Questions in 5 Major Tests," Reading Research Quarterly, 9, No.2:206-223, 1973-74.

VARLAAM, A., "Educational Attainment and Behavior at School," Greater London Intelligence Quarterly, No. 29:29-37, December, 1974.

WEINTRAUB, SAMUEL et al, "Summary of Investigations Relating to Reading July 1, 1973 to June 30, 1974," Reading Research Quarterly, 10, No.3:267-543, 1974-75.

WEINTRAUB, SAMUEL et al, "Summary of Investigations Relating to Reading, July 1, 1974 to June 30, 1975," Reading Research Quarterly, 11:223-563, 1975-76.

WELTON, MICHAEL, "The New Illiteracy: A Critical Response," Community Schools, 4, No.3:20-22, April, 1974.

BIBLIOGRAPHY FOR WRITING ASSESSMENT

BULLOCK, SIR ALAN. A Language for Life: Report of the Committee of Inquiry Appointed by the Secretary of State for Education and Science Under the Chairmanship of Sir Alan Bullock F. B. A. London: Her Majesty's Stationery Office, 1975.

FOLEY, JOSEPH J., "Evaluation of Learning in Writing," Chapter 21 in Handbook on Formative and Summative Evaluation of Student Learning, New York, N. Y.: McGraw-Hill Inc., 1971.

FORBES, RAY H., (Director). Writing Mechanics, 1969-1974: A Capsule of Changes in Writing Mechanics. (NAEP). Washington, D. C.: U. S. Government Printing Office, October, 1975.

FORBES, RAY H., (Director). Writing: Released Exercise Set (1973-74). (NAEP). Washington, D. C.: U. S. Government Printing Office, May, 1976.

HAZLETT, JAMES A., (Director). 1969-1970 Writing: National Results. (NAEP). Washington, D. C.: U. S. Government Printing Office, November, 1970.

MAXWELL, JOHN C., "National Assessment of Writing: Useless and Interesting?," English Journal, 62, No.9:1249-1253, December, 1973.

MOORE, WALTER J. and LARRY D. KENNEDY, "Evaluation of Learning in the Language Arts," Chapter 15 in the Handbook on Formative and Summative Evaluation of Student Learning, New York, N. Y.: McGraw-Hill Inc., 1971.

MULLIS, INA V. Highlights and Trends from National Assessment: Writing and Change in Writing Skills. San Francisco: American Educational Research Association, April, 1975. 25pp. (Unpublished Paper).

MULLIS, INA V. The Primary Trait System for Scoring Writing Tasks. (NAEP). Denver, Colorado, May, 1976. 32pp. (Unpublished Paper).

National Assessment of Educational Progress: Writing Objectives for 1973-1974 Assessment. Denver, Colorado: National Assessment Office, 1972. 33pp.

"National Assessment of Educational Progress: Teenagers' Writing Skills Decline," NAEP Newsletter, VIII, No.6:1-3, December, 1975.

Sample of a Local Edmonton School District Application of the Assessment of Writing Using the Primary Trait System. Edmonton, Alberta: Edmonton Public School District, May, 1976. (Unpublished Report).

SLOTNICK, HENRY B., and JOHN V. KNAPP, "Essay Grading by Computer: A Laboratory Phenomenon?," English Journal, 60, No.1:75-87, January, 1971.

SLOTNICK, HENRY B., (Assistant Director). Writing: National Results-Writing Mechanics. (NAEP). Washington, D. C.: U. S. Government Printing Office, February, 1972.

SLOTNICK, HENRY B., (Assistant Director). Selected Essays and Letters: A Selection of Papers Collected During the 1969-1970 Assessment of Writing. (NAEP). Washington, D. C.: U. S. Government Printing Office, November, 1972.

SLOTNICK, HENRY B., "Toward a Theory of Computer Essay Grading," Journal of Educational Measurement, 9, No.4:253-263, Winter, 1972.

SLOTNICK, HENRY B., "Close Up: The National Assessment of Writing Skills on the Teaching of Writing. Some Implications from National Assessment," English Journal, 62, No.9:1249-1253, December, 1973.

BIBLIOGRAPHY FOR FUNCTIONAL LITERACY ASSESSMENT

ALLEN, GEOFFREY C., "Provision for Adult Illiteracy in England and Wales," The Vocational Aspect of Education, XXVII, No.66:27-35, Spring, 1975.

AYRER, JAMES E. Development of A Test of Functional Literacy. San Francisco: American Educational Research Association, April, 1976. 12pp. (Mimeographed).

CAUGHRAN, ALEX M., and JOHN LINDLOF, "Should the 'Survival Literacy Study' Survive?," Journal of Reading, 15, No.6:429-435, March, 1972.

COTTERELL, ARTHUR, "Functional Literacy: A World Perspective on Reading," The Technical Journal, 11:18-23, October, 1973.

CULLEN, M. and D. PICKERING, "Breakthrough to Literacy," Curriculum and Research Bulletin, 6:42-46, June, 1971.

DOWNING, JOHN, "Cultural Priorities and the Acquisition of Literacy," International Review of Education, 19, No.3:345-354, 1973.

Functional Literacy - Basic Reading Performance an Assessment of In-School 17-year-olds in 1974. Denver, Colorado: Education Commission of the States, 1975. 20pp. (Press Summary). ED112389.

GADWAY, CHARLES and H. A. WILSON. Functional Literacy: Technical Summary. (NAEP). Washington, D. C.: U. S. Government Printing Office, 1975. 33pp.

GRANT, W. VANCE, "Illiteracy in the United States," American Education, 7, No.33, June, 1971.

Guidelines for Refining Existing Reading Programs to Ensure Teaching of a Survival Component. Maryland: State Department of Education, 1973. 16pp. (Mimeographed).

HAEMMERLI, A., "Some Observations of the Methodological Nature Concerning the Evaluation of the Experimental World Literacy Programms," Convergence, 5, No.1:66-69, 1972.

HANSEN, LEE H., and KARL D. HESSE. Criterion-Referenced Assessment of Reading Literacy Using the Cloze Procedure. Chicago, Illinois: Annual Meeting of the American Educational Research Association, April, 1974. 17pp. (Unpublished Paper).

HARMAN, DAVID, "Illiteracy: An Overview," Harvard Educational Review, 40, No.2:226-264, May, 1970.

HARRIS, LOUIS. Survival Literacy Study. New York: National Reading Council, September, 1970. 38pp. ED068813.

HARRIS, LOUIS. The 1971 Reading Difficulty Index: A Study of Functional Reading Ability in the U. S. New York, 1971. ED057312.

HILLERICH, ROBERT L., "Toward an Assessable Definition of Literacy," English Journal, 65:50-55, February, 1976.

International Symposium for Literacy Final Report. Iran, September, 1975. 39pp. ED123369.

JACKSON, REX. The Development of a Collection of Adult Tasks for Assessment of Literacy. Chicago, Illinois: American Educational Research Association, April, 1972. (Unpublished Paper). ED061023.

LICHTMAN, MARILYN. The Development and Validation of R/EAL, An Instrument to Assess Functional Literacy. Washington, D. C.: The Catholic University of America, June, 1973. 37pp.

MAHER, N. J., "Illiteracy," Educational Magazine, 29, No.9:38-39, 1972.

MASON, HORACE, "Fundamental Education and Functional Literacy-Problems and Possibilities," Convergence, VI, Nos.3 and 4:55-61, 1973.

McDONALD, THOMAS, F. and GARY B. MOORMAN, "Criterion-Referenced Testing for Functional Literacy," Journal of Reading, 17, No.5:363-366, February, 1974.

MURPHY, RICHARD T. Adult Functional Reading Study. Princeton: Educational Testing Service, December, 1973a. 113pp.

MURPHY, RICHARD T. Adult Functional Reading Study, APPENDIX A. Princeton: Educational Testing Service, December, 1973b. 69pp.

MURPHY, RICHARD T. Adult Functional Reading Study, APPENDIX B and APPENDIX C. Princeton: Educational Testing Service, December, 1973c. 236pp.

MURPHY, RICHARD T. Adult Functional Reading Study, Supplement. Princeton: Educational Testing Service, January, 1975. 130pp.

NAFZIGER, DEAN H., BRENT R. THOMPSON and MICHAEL D. HISCOX. Tests of Functional Adult Literacy: An Evaluation of Currently Available Instruments. Portland: Northwest Regional Educational Laboratory, January, 1976. 110pp.

National Assessment of Educational Progress. "17-Year-Olds Gain in Basic Reading Skills 'Functional Literacy' Study Shows," NAEP Newsletter, VIII, No.5:105, October, 1975.

NORTHCUTT, NORVELL. Functional Literacy for Adults: a Status Report of the Adult Performance Level Study. New Orleans: International Reading Association, May, 1974. 14pp. (Unpublished Paper). ED091672.

STICHT, THOMAS G. et al, "Project Realistic: Determination of Adult Functional Literacy Skill Levels," Reading Research Quarterly, 7, No.3:424-465, Spring, 1972.

THOMAS, AUDREY M. Adult Basic Education and Literacy Activities in Canada 1975-76. Toronto: World Literacy of Canada, April, 1976. 180pp.

"The World Problem of Illiteracy," NEA Research Bulletin, 50, No.2:53-58, May, 1972.

BIBLIOGRAPHY FOR ORAL COMMUNICATION ASSESSMENT

BULLOCK, SIR ALAN. A Language for Life. Report of the Committee of Inquiry Appointed by the Secretary of State for Education and Science Under the Chairmanship of Sir Alan Bullock F. B. A. London: Her Majesty's Stationery Office, 1975.

CARROW, ELIZABETH, "A Test Using Elicited Imitations in Assessing Grammatical Structure in Children," Journal of Speech and Hearing Disorders, 39, No.4:437-444, November, 1974.

CAZDEN, COURTNEY B., "Hypercorrection in Test Responses," Theory into Practise, XIV, No.5:343-346, December, 1975.

DEYES, A. E., "How Well Can Our Students Speak? (Language Laboratory Testing)," English Language Teacher, 26, No.31-7, October, 1971.

DOBBYN, MICHAEL, "An Objective Test of Pronunciation for Large Classes," English Language Teaching Journal, 30:242-244, April, 1976.

FOLLAND, DAVID and DAVID ROBERSTON, "Towards Objectivity in Group Oral Testing," English Language Teacher, 30:156-167, January, 1976.

GRAHAM, NORMAN, "Towards a Test of Language Production," Educational Review, 24:34-46. November, 1971.

LEVINE, JOSIE, "An Outline Proposal for Testing Communicative Competence," English Language Teacher, 30:128-35, January, 1976.

MAXWELL, J. M. Vocabulary Usage and the Classification of Primary Children. New Orleans, Louisiana: National Council of Measurement in Education, February, 1973. (Unpublished Paper).

MELEAR, JOHN D., "An Informal Language Inventory," Elementary English, 51:508-511, April, 1974.

MOORE, WALTER J., and LARRY D. KENNEDY, "Evaluation of Learning in the Language Arts," Chapter 15 in Handbook on Formative and Summative Evaluation of Student Learning, McGraw-Hill Inc., New York, 1971.

PALMER, ADRIAN S., "Testing Communication," IRAL, 10:35-45, February, 1972.

POTTER, ROBERT E. and DAVID B. STROTHER, "A Means of Accountability in Oral Communication Performance," Speech Teacher, 24:146-150, March, 1975.

ROBINSON, PETER, "Oral Expression Tests," English Language Teaching, 25:151-5, 260-66, February, 1971.

ROSE, SUZANNA, et al. The Development of a Measure to Evaluate Language Communication Skills of Young Children. New Orleans, Louisiana: American Educational Research Association, March, 1973. (Unpublished Paper).

SEWARD, B. H., "Measuring Oral Production in EFL," English Teaching Journal, 28:76-80, November, 1973.

SKULL, J., and A. WILKINSON, "The Construction of an Oral Composition Quality Scale," British Journal of Educational Psychology, 39:272-277, November, 1969.

STAUFFER, RUSSELL G., and JOHN J. PIKULSKI, "A Comparison and Measure of Oral Language Growth," Elementary English, 51, No.8:1151-1155, Nov/Dec, 1974.

TUCKER, CHARLES O., "Testing and Evaluative Procedures in Speech Communication," The Bulletin of NASSP, 54, No.350:114-122, December, 1970.

WILLIAMS, MAURICE, EDITH WEINSTEIN and RALPH O. BLACKWOOD, "An Analysis of Oral Language Compared With Reading Achievement," Elementary English, XLVII, No.3:394-396, March, 1970.

WILKINSON, A., and LESLIE STRATTA, "The Evaluation of Spoken Language," Educational Review, 21:183-195, June, 1969.

WILKINSON, A., "The Quality of Language Experiences in Younger Children," Journal of Curriculum Studies, 240-246, November, 1969.

BIBLIOGRAPHY FOR LISTENING ASSESSMENT

BOWDIDGE, JOHN S. Cloze Listening Test (Form Lisbon and Form Waco). Springfield, Missouri: Drury College, 1967. (Unpublished Manuscript).

BULLOCK, SIR ALAN. A Language for Life. Report of the Committee of Inquiry Appointed by the Secretary of State for Education and Science Under the Chairmanship of Sir Alan Bullock F.B.A. London: Her Majesty's Stationery Office, 1975.

BUTLER, JOSEPH. Comparisons of Listening Abilities, Categorized as Good and Poor, of Inner-City Children in the Sixth Grade. Indiana: Indiana University, 1970. 128pp. (Unpublished Paper). ED054177.

CARVER, RONALD P., "Learning From Prose Materials," Journal of Educational Psychology, 64:76-84, 1973.

FINKENBINDER, RONALD L., "A Descriptive Study of the Goldman-Fristoe-Woodcock Test of Auditory Discrimination and Selected Reading Variables with Primary School Children," The Journal of Special Education, Vol.7, No.2:125-131, 1973.

GOLDMAN, RONALD, MACALYNE FRISTOE, RICHARD WOODCOCK, "A New Dimension in the Assessment of Speech Sound Discrimination," Journal of Learning Disabilities, 4, No.7:364-368, Aug/Sept., 1971.

HACKETT, MARIE GANNON. A Hierarchy of Skills in Listening Comprehension and Reading Comprehension. Tallahassee: Florida State University, Department of Educational Research, 1970. 27pp. (Unpublished Manuscript).

KENNEDY, DELORES KESSLER and PAUL WEENER, "Visual and Auditory Training with the Cloze Procedure to Improve Reading and Listening Comprehension," Reading Research Quarterly, 8:524-65, Summer, 1973.

MECHAM, MERLIN J., "Measurement of Verbal Listening Accuracy in Children," Journal of Learning Disabilities, 4, No.5:27-29, May, 1971.

MOORE, WALTER J. and LARRY D. KENNEDY, "Evaluation of Learning in the Language Arts," Chapter 15 in Handbook on Formative and Summative Evaluation of Student Learning, New York: McGraw-Hill, Inc., 1971.

NESBITT, MARY CATHERINE. Auding Achievement of First Grade Pupils Related to Selected Pupil Characteristics. Georgia: University of Georgia, 1968. 112pp. (Unpublished Manuscript). ED034785.

PALMATIER, ROBERT A. and GEORGE McNINCH, "Source of Gains in Listening Skill: Experimental or Pre-Test Experience?," The Journal of Communication, Vol. 22:70-76, March, 1972.

PENFIELD, D. A. and L. A. MARASCUILO, "Learning to Listen: An Experimental Study", Educational Research, 14:220-4, June, 1972.

PRICE, JACK ALBERT. The Influence of Three Listening-Environment Factors on Listening Comprehension of Fourth-Grade Students. Texas: North Texas State University, 1969. 134pp. (Unpublished Manuscript). ED048263.

ROSNER, JEROME and DOROTHEA SIMON, "The Auditory Analysis Test: An Initial Report," Journal of Learning Disabilities, 4, No.7:384-392, Aug/Sept., 1971.

ROSNER, JEROME, "Auditory Analysis Training with Prereaders," The Reading Teacher, 27, No.4:379-384, January, 1974.

SCHWARTZ, ARTHUR H. and RONALD GOLDMAN, "Variables Influencing Performance on Speech-Sound Discrimination Tests," Journal of Speech and Hearing Research, 17:25-32, 1974.

WALLNER, NANCY KUBIN, "The Development of a Listening Comprehension Test for Kindergarten and Beginning First Grade," Educational and Psychological Measurement, 34:391-396, Summer, 1974.

WEAVER, SUSAN and WILLIAM L. RUTHERFORD, "A Hierarchy of Listening Skills," Elementary English, 51, No.8:1146-1150, Nov/Dec., 1974.

WILKINSON, ANDREW, "Research in Listening Comprehension," Educational Research, 12:140-144, No.3, 1970.

WILKINSON, ANDREW and LESLIE STRATTA, "Listening Comprehension at 13 Plus," Educational Review, 22:228-242, June, 1970.

WILKINSON, ANDREW, LESLIE STRATTA and PETER DUDLEY. The Quality of Listening. London: Macmillan Education Ltd., 1974.

WILLIAMS, GEORGE L., "Self-Evaluation in Speaking and Listening," English Journal, 61, No.1:68-70, January, 1972.

APPENDIX A:

SUMMARIES OF RESEARCH RELATED TO ASSESSMENT OF READING COMPREHENSION

EVALUATING PERFORMANCE ON NATIONAL ASSESSMENT OBJECTIVES: NORM-REFERENCED CRITERION-REFERENCED INTERPRETATIONS

David J. Alvord
Barbara E. Brittingham

Iowa
1974

Purpose

The purpose of this study was to utilize National Assessment of Educational Progress (NAEP) exercises to assess performance on specific reading objectives. The method of data analysis was to encompass both norm-referenced and criterion-referenced interpretations of pupil performance.

Methodology and instrumentation

Forty-six exercises from NAEP were administered to a stratified random sample of 1,105 Iowa fourth grade students. Directions and procedures for administration were provided in script form and pupils were assessed in groups no larger than ten to twelve students

To determine whether the performance on exercises related to an objective was sufficient to suggest attainment of the objective, two approaches were used.

1. A norm-referenced approach in which the performance of Iowa pupils was compared to a norm group of 9-year-olds who had written the same exercises in 1970-71.

2. A criterion-referenced approach in which performance was compared to the estimated difficulty levels originally assigned to the exercises by NAEP (easy-difficulty = .90, medium-difficulty = .50, and hard-difficulty = .10).

Analysis

The Wilcoxson Signed Ranks Test was used to test for differences between the median performance of Iowa pupils on exercises related to a given objective and (1) the median performance of the norming group on the same exercises and (2) the median estimated difficulty level for those exercises as reported by NAEP.

Major outcomes and interpretations

1. From both viewpoints, the performance of Iowa pupils on all reading objectives was equal to or above the standards set by NAEP and the norm group.

2. Certain problems related to the use of the NAEP estimated difficulty levels in the criterion-referenced aspect of the analysis were identified.

3. The study suggests the following factors be considered when setting criterion levels for these tasks:

   (a) the taxonomic level of the objective;

   (b) the number of words to be read in the item;

   (c) the number of words contained in any material preceding the exercise; and

   (d) the kinds of illustrations or graphs associated with the exercise.

HOW TO CONSTRUCT ACHIEVEMENT TESTS TO ASSESS COMPREHENSION

Richard C. Anderson Illinois
1972

Purpose

The purpose of this project was to establish the necessary distinctions to indicate whether or not a person had comprehended an instructional communication, to outline procedures for constructing test questions based on these distinctions, and to show that these test questions gave rise to orderly, sensible data with respect to reading comprehension.

Report content

Anderson contended that it was only necessary to distinguish between comprehension and noncomprehension. Many other test developers usually considered several finer distinctions when developing test questions. Anderson first discussed two item types which did not require comprehension skills to be answered correctly. These types were verbatim questions and transformed or rearranged verbatim questions. He then stated that comprehension could be inferred from the capacity to answer a question based on paraphrase of a test statement. The following definition further explained the paraphrase question.

Two statements were defined as paraphrases if (1) they had no nouns, verbs, or modifiers in common and (2) they were equivalent in meaning. Test questions were developed from these paraphrases by deleting an element of the paraphrase or changing a part of the paraphrase into a question. Paraphrase questions or transformed paraphrase questions were two types of questions formed in this manner. Comprehension of both concepts and principles were said to be assessable by the paraphrase procedure. Either the question was formed by paraphrasing the concept or principle itself, or the concept or principle was referred to and the student was then asked to supply a new application. The following is an example of a paraphrase question assessing comprehension of a concept.

1.	Original sentence	The bullet struck the bungalow.
2.	Valid paraphrase made by the test developer but not seen by the student.	The object hit the house.
3.	Valid paraphrase question asked of student on original sentence.	What hit the house?

4. Answer bullet

If the verbatim question "What struck the bungalow?" was used, the item would not be assessing comprehension according to Anderson.

Anderson further indicated that when instruction consisted of a series of instances of a principle, a student's comprehension could be tested by his capacity to "induce" the principle from the specific instances. This indicated his technique could be appropriate for measuring beyond literal comprehension.

Major outcomes and interpretations

1. Anderson, after reviewing the current status of achievement tests with respect to measurement of comprehension, concluded that few researchers have indicated any rationale for test question selection.

2. He contended that whether a test question measures comprehension depends upon the relationship of the wording of the test question to the wording of the instructions. He concluded that items based on paraphrases rather than verbatim recall or recognition questions would validly measure comprehension.

3. Anderson also claimed that the development of paraphrase items was reproducible from study to study although no definitive rules were presented in the report.

LEA-CLOZE COMPREHENSION TEST

Ralph Balyeat
Douglas Norman

Tennessee
1975

Purpose

The major purpose was to develop and test a LEA-cloze test as a predictor of reading achievement and as a method of classroom instruction. The LEA-cloze technique involves the application of the lexical cloze multiple-choice format to materials written or identified by children (thus the term Language Experience Approach). Specifically, the objectives were to provide a reliable indicator of reading achievement (1) for teachers to use in forming initial reading groups at the beginning of the year, (2) which could be administered in one sitting and be easily scored and interpreted by teachers, and (3) which would relate to the language and experience of rural mountain children.

Methodology and instrumentation

Materials suitable for grades one through nine were collected from thirteen counties throughout the region and the lexical cloze procedure was applied to the materials which were written or dictated by children. In this cloze format, nouns, main verbs and adjectives were identified. Then every fifth word which was one of these parts of speech was deleted. These five words were paired with distractor words of the same grammatical class and the ten words were then added in random order just below the line of copy in which the fifth deletion appeared.

In 1972-73, exercises at grades one to four levels were administered to grade two students while grade five students wrote levels one through six. One hundred twenty students wrote the first year. In the following year ninety-two of the same students wrote levels one through six and four through nine. Pupils were asked to complete as many exercises as they could and there was no time limit.

Analysis

LEA-cloze scores from tests given in March, 1973 and March, 1974 were correlated with scores on the Stanford Achievement Test (SAT) reading-related subtests administered in May, 1973 and a t-test was applied.

Major outcomes and applications

1. The study found a strong positive correlation between LEA-cloze and SAT.

2. Since pupils in effect write or otherwise provide the materials on which tests are built, the LEA-cloze may hold special promise for measuring reading comprehension of culturally deprived children.

Sample

The coach scolded Randy who was playing the _____ of pitcher on the Little League team.
1

"Why do you refuse to keep your _____ on the ball?" the coach asked him.
2

Randy _____ up his nose. He looked disturbed but wouldn't _____.
3 4

"I am very disappointed in you," the coach told him. "You're not _____ a very tight game.
5

playing	position	wrinkled	foot	answer
panting	field	rolled	eye	carry

(Balyeat and Norman, 1975, p.557)

Figure 1
Sample Lexical Cloze Test

READING LITERACY: ITS DEFINITION AND ASSESSMENT

John R. Bormuth Illinois
1974

Purpose

This article had the following three purposes:

1. to analyze the concept of literacy for the purpose of identifying the parameters that must be specified in literacy definitions;

2. to identify measurement problems associated with each of these parameters; and

3. to describe literacy assessment procedures currently available for dealing with these measurement problems.

Report Content

The five parameters discussed by Bormuth were:

1. the classes of literacy behaviors;

2. the level of performance that serves as a criterion of literate performance;

3. the kinds of reading tasks on which the behaviors are tested;

4. the proportion of the reading tasks that serves as the criterion of literacy on some set of reading tasks; and

5. certain characteristics of the people to be tested.

In discussing these five parameters Bormuth attempted to define more closely what it means to be literate.

The first parameter dealt with the definition and assessment of literacy. Literacy behaviors were defined to include several skill areas, namely decoding, literal comprehension, inferential, critical reading,aesthetic appreciation, reading flexibility and study skills. Bormuth also discussed the reasons for only measuring overt behaviors indicative of reading literacy and the selection of a testing mode to measure these behaviors. This discussion was followed by an introduction of two procedures which would produce populations of items in a replicable manner using algorithms. The first or basic cloze procedure deleted specified single words in a passage and then

replaced them with a blank of standard length. This basic cloze procedure was limited to the measuring of decoding and literal comprehension since only single words were deleted. The second procedure, an adaptation of the basic cloze technique involved the development of "wh-questions" and was directly related to Osgood's (1963) theory of comprehension. This theory stated that comprehension took place by a series of events through which the meaning of one word or phrase was combined with or modified by the meaning of another word or phrase. Wh-questions or who, what kind, when, where, what, and why questions did allow for the measurement of inferential and critical reading behaviors in the project. Also included were the means to measure the ability to get the important points or main ideas from a passage and the ability to comprehend the structure of the material. Thus, this first parameter defined behaviors included in literacy and suggested procedures for measuring these behaviors.

The second parameter discussed was the level of performance a person should exhibit on a test of literacy before he could be regarded as literate. The operational problem was defined as one of assigning values to test scores, then identifying the score which had the greatest value. Value in this instance meant the amount of benefit associated with obtaining any given test score. Both positive and negative aspects were considered in making the decision. Thus, Bormuth's criterion identification model was based on the assumption that people read because doing so provided several different effects on them, each with its own value. Examples of possible effects were information gain, change in rate of reading, stimulation by difficult concepts, and interest in study. Different amounts of each effect could be associated with each cloze score and a value assigned to the effect. Each effect was then multiplied by its value. For each cloze score, the values of the individual effects were then summed. The cloze score which had the greatest value was designated the criterion score, that is, the score having the greatest overall benefit to the person.

Bormuth then reviewed the kinds of tasks on which the reading behaviors were to be tested. This formed his third parameter. The following five criteria were listed as being relevant to decisions about whether to include or exclude a set of reading tasks: monetary cost, economy of time, value or relative importance, commonness, and frequency of occurrence of the task in ordinary reading situations.

The fourth parameter was the proportion of reading tasks that should serve as the criterion of literacy. It was decided that the criterion for any set of reading tasks should be determined by using the criterion identification model previously discussed as the second parameter. Both negative and positive effects were associated with being able to read portions of the tasks at different identifiable levels. These effects could include such things as estimated

expected income, costs of instruction and time needed for instruction. These effects would be assigned values and the data would be entered into the model to find the optimal test score.

In the fifth parameter, Bormuth identified three characteristics of individuals which seemed relevant in identifying a literacy criterion score. These factors were native capacity to learn, acquired capacity to learn, and motivation. Considerations of these characteristics was incorporated into the criterion identification model in order to take the relevant major individual characteristics into account along with the other variables important in determining a criterion score of literacy.

Major outcomes and interpretations

In summary, Bormuth has developed an operational definition of and model for assessing reading literacy based on the assumption that a person's literacy is jointly determined by both his reading ability and the readability of the materials he needs to read.

The criterion identification model regarded a person as literate when he could perform well enough to obtain the maximum value from the materials he needed to read.

Bormuth further suggested that a criterion test which could successfully be used in the model could consist of a number of cloze or wh-questions. Both of these question types could be produced by an algorithm. Use of these question types would greatly increase the degree to which inferential and critical comprehension skills could be validly assessed in reading.

MEASURING THE PRIMARY EFFECT OF READING: READING-STORAGE TECHNIQUE, UNDERSTANDING JUDGMENTS, AND CLOZE

Ronald P. Carver Maryland
1974b

Purpose

The purpose of this study was to evaluate the newly develcped reading-storage (RS) test (Carver 1970), by comparing it to a modified version of the cloze (MC) test. The RS test was a completely objective type of cloze procedure which was produced from prose passages using algorithms. The three experiments involved the three different conditions of writing items after reading passages versus the usual writing of items without reading complete passages, varying degrees of passage deletion and varying the amounts of interpolated activity between passage and tests.

Methodology and instrumentation

Subjects were forty-eight students from the University of Maryland who read one hundred word passages (one minute time limit) at four levels of difficulty, then answered either a RS or MC test immediately.

In a reading-storage item, every other word in the one hundred word paragraph was deleted except for the initial letter. The remaining words and blanks were separated into ten lines with five blanks per line. One of the five initial letters was replaced with a wrong letter. The task for the individual was to read the passage and then recognize and circle the wrong letter on each line without referring back to the passage.

In the MC tests constructed, every fourth, fifth, and sixth word was deleted and replaced with the initial letter of the word. The task for the subject was to fill in the blanks with the correct words using the initial letters as clues. The time limit was seven minutes.

Examples of RS and MC tests may be found at the end of this summary.

Analysis

The following comparisons were made:

1. understanding as a function of difficulty level for the RS test groups and MC test groups;

2. understanding as a function of the percentage of passage deletion for RS test groups and MC test groups;

3. test score as a function of difficulty level for RS test groups and MC test groups under reading and non-reading conditions;

4. test score as a function of percent of passage deletion for RS and MC test structures; and

5. test score as a function of the amount of interpolated activity for RS and MC test structures and a combination of the two structures.

Major outcomes and interpretations

1. Both tests seemed to reflect about the same amount of average gain from the reading to the non-reading condition. That is, scores in both tests improved when the passage was read prior to the test being written.

2. Both the RS test and the MC test appeared to be sensitive to decrements in the amount of information presented--the more information presented, the higher the scores in both tests.

3. The RS test appeared to reflect a higher amount of information stored during the normal reading condition than the MC test. Since scores did not drop sharply with interpolated activity between the passage and the RS test, the RS test would be useful as a measure of reading comprehension.

Samples

This is our Post Office. It is in our city. Many people work here. There is a Post Office in every city in our country. And Post Offices in every country in the world.

A Post Office helper must be honest. He must be a good worker. A Post Office helper handles lots of mail. A Post Office helper handles lots of money.

The Post Office sends letters and packages, magazines, and newspapers all over the world. It sends small animals and plants, too. It sends money for us. It saves money for us. It puts money to work for us, too.

(Carver, 1974b, p. 253)

Figure 2
Sample Passage on which Reading-Storage (RS) Test and Modified-Cloze (MC) Test are Based

THIS IS

1. O___ POST O___. IT I___ IN (H)___ CITY. M___ PEOPLE
2. W___ HERE. (O)___ IS A___ POST O___ IN E___ CITY
3. (L)___ OUR C___. AND P___ OFFICES I___ EVERY C___ IN
4. T___ WORLD. A___ POST O___ HELPER M___ BE (S)___. HE
5. M___ BE A___ GOOD W___. A (E)___ OFFICE H___ HANDLES
6. S___ OF M___. A P___ OFFICE H___ HANDLES L__ OF
7. M___. THE C___ OFFICE S___ LETTERS A___ PACKAGES, M___, AND
8. N___ ALL O___ THE W___. IT S___ SMALL W___ AND
9. H___, TOO. I___ SENDS M___ FOR U___. IT S___ MONEY
10. F__ US. A ___ P.S M___ TO W___ FOR U___, TOO.

(Carver, 1974b, p. 254)

Figure 3
Sample Exercise from a Reading-Storage (RS) Test for the Passage in Figure 2.

THIS IS OUR	P Post	O Office	I it
IS IN OUR	C city	M Many	P people
WORK HERE. THERE	I is	A a	P Post
OFFICE IN EVERY	C city	I in	O our
COUNTRY. AND POST	O Offic	I in	E every
COUNTRY IN THE	W world	A a	P Post
OFFICE HELPER MUST	B be	H honest	H He
MUST BE A	G good	W worker	A a
POST OFFICE HELPER	H helles	L lots	O of
MAIL. A POST	O_____	H_____	H_____
LOTS OF MONEY.	T_____	P_____	O_____
.....	.....	.....	.....

(Carner, 1974b, p. 157)

Figure 4
Sample Exercise from a Modified-Cloze (MC) Test for the Passage in Figure 2.

COMPARING THE READING-STORAGE TEST TO THE PARAPHRASE TEST AS MEASURES OF THE PRIMARY EFFECT OF PROSE READING

Ronald P. Carver Washington 1975a

Purpose

Two experiments were completed in this study. The primary purpose of Experiment 1 was to investigate the sensitivity of the reading-storage (RS) test to the primary effect of reading as compared to the paraphrase test (P-test) at four different levels of prose difficulty. The purpose of Experiment 2 was to compare the RS test and P-test under conditions where the amount of information presented was manipulated by deleting varying percentages of words from the reading passages.

Methodology and instrumentation

In Experiment 1, each of forty-eight college level paid subjects received both an RS-test and a P-test on eight one-hundred word passages at four levels of difficulty under reading and non-reading conditions. Each passage had a one-minute time limit and each test a three-minute time limit. The RS-test was exactly the same test as in the 1974 RS-cloze comparison study (Carver 1974b). The paraphrase questions were developed according to recommendations given by Anderson (1972) but modified by writing a five-choice multiple-choice item on every sentence in each passage. The question and correct alternative represented the valid paraphrase.

In Experiment 2, forty college level paid subjects received both an RS-test and a P-test under reading and non-reading conditions, at one level of difficulty but with five different levels of information presented. The RS-test was exactly as in the RS-cloze test used in the study conducted by Carver (1974b). In this study, five one-hundred word passages were subjected to a systematic word deletion pattern from zero to one hundred per cent. P-tests on the same five passages were developed under guidelines the same as given in Experiment 1 above.

Analysis

For each of Experiment 1 and Experiment 2 and as for the 1974 studies, the RS-test and P-test were compared to determine which test had the higher gains from reading to non-reading and which test had the higher efficiency ratio. An efficiency ratio was defined as the test score divided by the reader's subjective rating of the per cent of the passage he understood.

Major outcomes and applications

1. As passages became more difficult, fewer RS-test items could be answered correctly without reading.

2. Using the understanding ratings as a control in the efficiency ratio provided a good quantitative technique for measuring the primary effect of reading which is reading comprehension.

3. The RS-tests were less sensitive to gain from reading than the P-test but gain on the P-test was more variable.

4. The subjectively developed P-test seemed to have a higher potential (higher efficiency ratio) than the objectively developed RS-test, but in any particular situation, they were less sensitive to the primary effect of reading than the RS-test.

5. This study suggested that the RS-test could be more appropriate than the P-test when measurement of gain due to reading is the primary objective.

REVISED PROCEDURES FOR DEVELOPING READING-INPUT MATERIALS AND READING-STORAGE TESTS

Ronald P. Carver Washington 1975b

Purpose

This study had three major purposes. It provided a simplified algorithm for the development of reading-input (RI) materials. These materials were produced from prose passages using algorithms. A sample is given at the end of the summary. It provided the algorithm for developing the revised version of the reading-storage test (RS). A sample is given at the end of the summary. The study compared, according to nine criteria, the reading-input (RI) tests and reading-storage (RS) tests with multiple-choice (MC) tests and cloze (CLZ) tests.

Methodology and instrumentation

1. The revised steps for producing reading-input materials were presented.

2. The revised steps for developing reading-storage tests were presented.

Analysis

The four techniques, RI, RS, MC, and CLZ were discussed with respect to the nine criteria. These criteria were

1. objective scoring,
2. relationship between amount of prose and number of items,
3. objectivity in test development,
4. ease of item development,
5. task time limits and task difficulty,
6. investigating readability,
7. measuring the primary effect of reading,
8. inducing prose learning, and
9. increasing reading skill.

Major outcomes and interpretations

1. RI and RS tests were found to be completely objective in both scoring and question development.

2. RI tests seemed to encompass all the advantages of cloze, yet required less time for the examinee to complete. RI tests were the best measure of readability of the four tests. This meant they provided the best estimate of reading difficulty level for a prose passage.

3. RI tests also appeared to be the best measure of inducing prose learning and showed good potential as a method of increasing reading skill.

4. RS tests seemed to have many of the advantages of MC tests while overcoming many of the disadvantages of MC tests.

5. RS tests reflected score gain due to reading better than MC, RI or regular cloze tests and about as well as P-tests constructed from paraphrase questions. RS tests were also passage-dependent and thus the most appropriate of the four tests to measure the primary effect of reading, reading comprehension.

Samples

This is ☐ country ☐ our Post Office. It is ☐ there ☐ in our city.

Many people ☐ post ☐ work here. There is a ☐ Post ☐ Be Office in

every city ☐ is ☐ in our country. And Post ☐ Good ☐ Offices in

every country in ☐ the ☐ city world.

A Post Office ☐ helper ☐ in must be honest. He ☐ handles ☐ must

be a good worker. ☐ Every ☐ A Post Office helper handles

☐ lots ☐ and of mail. A Post ☐ animals ☐ Office helper handles lots of ☐ and. ☐ money.

The Post Office sends ☐ worker ☐ letters and packages, magazines,

and ☐ helper ☐ newspapers all over the world. ☐ It ☐ Helper sends small

animals and ☐ plants, ☐ all, too. It sends money ☐ for ☐ office us. It

saves money ☐ sends ☐ for us. It puts money ☐ to ☐ and work for us too.

(Carver, 1975b, p. 157)

Figure 5
A Sample of Applying the Reading-Input (RI) Procedures.

T____ IS
1. O____ POST O____. IT I____ IN H____ CITY. M____ PEOPLE
2. W____ HERE. O____ IS A____ POST O____ IN E____ CITY
3. I____ OUR C____. AND P____ OFFICES I____ EVERY C____ IN
4. T____ WORLD. A____ POST O____ HELPER M____ BE S____. HE
5. M____ BE A____ GOOD W____. A P____ OFFICE H____ HANDLES
6. S____ OF M____. A P____ OFFICE H____ HANDLES L____ OF
7. M____. THE C____ OFFICE S____ LETTERS A____ PACKAGES, M____, AND
8. N____ ALL O____ THE W____. IT S____ SMALL W____ AND
9. H____, TOO. I____ SENDS M____ FOR U____. IT S____ MONEY
10. F____ US. A____ PUTS M____ TO W____ FOR U____, TOO.

(Carver, 1975b, p. 165)

Figure 6
A Sample Reading-Storage (RS) Test.

FURTHER RESEARCH ON THE READING-STORAGE TEST AS A MEASURE OF GAIN DURING READING

Ronald P. Carver Washington 1975c

Purpose

This article, adding to the Bormuth Cloze Technique, was related to the passage dependency problem. Scores on some standardized reading tests are more dependent upon reading the passage than other tests. A modified cloze technique called the Reading-Storage (RS) technique was investigated to determine its sensitivity to gains in score due to reading of the passages ahead of writing the RS test items. Gains were also related to passage difficulty, reader ability and the readers' opinions of how well they understood the passage.

Methodology and instrumentation

The six experimental passages were sampled from the 330 one hundred word passages developed by Bormuth (1974). These passages, when rated for difficulty level according to two different readability scales, rated from beginning reader level to college level. The RS tests were administered to 600 students in grades four to twelve in a small rural school system. About 400 of these students had also written the National Reading Standards (NRS) Test and thus two estimates of reader ability were available.

After reading a passage, the students judged the degree to which they understood the passage and were administered an RS test on the passage. Each subject wrote six tests of the same difficulty level: tests 1, 3, and 5 were written after reading the passage, while tests 2, 4, and 6 were administered without any opportunity to read the passage on which the tests were based.

Analysis

Each RS test was scored, a correction for guessing formula was applied, and then scores were converted into a percent correct score. For all students at each level of reader ability, the mean of the RS test scores on all the passages at each passage difficulty level was calculated. The same type of analysis was used for the "percent understanding judgments" except medians were used.

Major outcomes and interpretations

1. Gain scores on the RS test tended to be higher for low difficulty material than high difficulty material. That is, scores between reading and non-reading conditions differed more for easy material than for difficult material.

2. Gain scores on the RS test tended to be higher for high ability individuals than low ability individuals.

3. There seemed to be a high relationship between gains on the RS test and readers' understanding judgments. This would indicate that an RS test could be used to indicate the degree to which a passage was understood by the readers.

4. The RS test appeared to be reasonably sensitive to gain due to reading as long as the difficulty level of the material was not more than one level above or below the ability of the individual writing the test.

5. Since tests administered at various grade levels were approximately equal in difficulty to the average ability of students at that grade level it appeared that the RS test would be a sensitive measure to assess gain in score due to reading. This would provide a desired measure of reading comprehension.

A QUESTION WRITING ALGORITHM

Patrick J. Finn

New York
1975

Purpose

A major purpose of this study was to generate and test additional rules and a subsequent algorithm to develop test items and ensure that independent item writers would generate the same legitimate achievement test questions from any passage. The algorithm consisted of rules for analysis of sentences in preparation for item writing and question writing rules. This study carried on earlier work completed by Bormuth (1974) who generated questions by rules which refer to the syntax. A tree diagram drawn over the surface sentences labelled each node S (structural) or L (lexical) and generated Wh-rote questions by randomly deleting lexical nodes.

Methodology and Instrumentation

Five linguistics students at the University of Edinburgh volunteered to be paid subjects in an experiment to develop and test the algorithm. Very briefly the development of the algorithm could be described as follows. The syntactic analysis used was based on the Basic Rules of Integration of Transformational Theories of English Syntax (Stockwell et al, 1968). Question writing rules were continuously developed and revised by the subjects as different sentences underwent the analysis and item writing procedures. The final version consisted of eighty-two question-writing rules.

The algorithm was tested by selecting three 500-word passages. Each of the five item writers generated twenty questions yielding an opportunity to test one hundred questions for conformity. In order for all five item writers to have written the same question from the passage they must have chosen the same sentence, deleted the same question node, chosen the same Wh-word, included other clauses according to the algorithm, and employed the same word order. It was hypothesized that proper use of the algorithm would lead to conformity in all areas. Examples of non-conformity were listed.

Analysis

After all questions were generated the number of trials conforming to the algorithm were tabulated. These were listed on the basis of adherence to the five criteria for conformity just mentioned above.

Major outcomes and interpretations

1. Item writers chose the same sentences ninety-seven times out of one hundred, the same node to questions ninety-two

times out of one hundred, the same node and exactly the same words in the remainder of the sentence eighty-four times out of one hundred. Seventy-seven times out of one hundred they produced a question with exactly the same word order.

2. This algorithm enabled writers to objectively generate questions by definition and at the same time not produce anomalous questions. The use of the algorithm did not generate all questions possible, but it did vigorously define the universe of questions that could be generated.

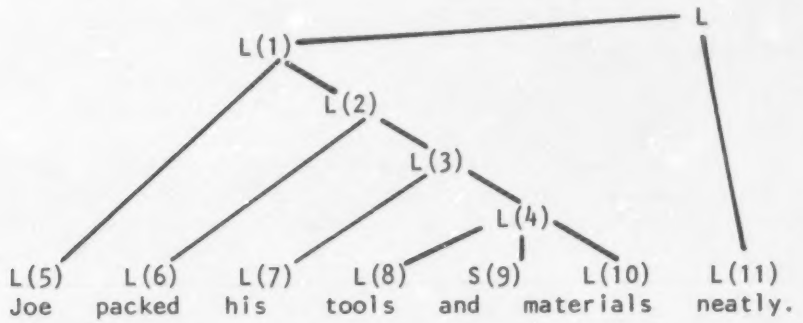
3. The development and testing of this algorithm for generating test items by definition from any passage should benefit objective test construction in the area of reading comprehension and make test scores more meaningful.

Samples

Bormuth's Proposal

Bormuth proposed generating questions from written texts by analyzing the syntax of the text and generating questions by rules which refer to the syntax. One draws a tree diagram over the surface sentences, labels each node S (structural) or L (lexical), and generates Wh-rote questions by randomly deleting lexical nodes. For example, questions in Figure 7 are produced from sentence (1).

(1) Joe packed his tools and materials neatly.

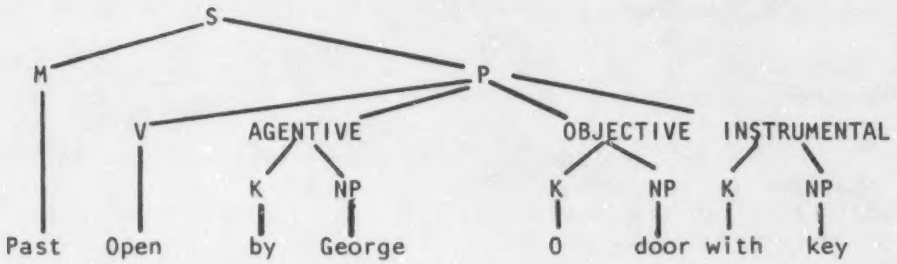


Questions Generated by Deleting Odd Numbered Nodes in Sentence (1)

Nodes Deleted	Question Generated
4	What did Joe pack neatly?
5	Who packed his tools and materials neatly?
7	Whose tools and materials did Joe pack neatly?
9	No question: Structure words are not deleted.
11	How did Joe pack his tools and materials?

(Finn, 1975, p. 342)

Figure 7
The Use of Syntactic Analysis as a Vehicle for Explicitly Specified Processes of Test Construction.



POSSIBLE SURFACE STRUCTURE SENTENCES

1. George opened the door with a key.
2. The door was opened by George with a key.

POSSIBLE Wh-QUESTIONS

DELETE:	GENERATE
VERB	What did George do to the door with a key?
AGENTIVE	Who opened the door with a key?
OBJECTIVE	What did George open with a key?
INSTRUMENTAL	With what did George open the door?

(Finn, 1975, p. 345)

Figure 8
Underlying Structure for the Bormuth Model - Details Omitted.

PREDICTING TWELFTH GRADERS' COMPREHENSION SCORES

Albert J. Harris
Milton D. Jacobson

Virginia
1976

Purpose

The major purpose of this study was to develop a readability formula which estimated the percentage of comprehension questions that high school seniors would be able to answer after reading a passage.

Methodology and instrumentation

A sample of 3,000 seniors from sixty school districts in Virginia was selected as representative of a nationwide sample on the basis of their SCAT and STEP scores. These students wrote five sets of standardized multiple-choice tests (McCall-Crabbs Standardized Test Lessons in Reading, 1961). Since the mean correct responses for the thirty-two passages was 76.8% and the standard deviation equalled 7.9 the passages appeared to be at a satisfactory level of difficulty for the population.

Analysis

Each of the thirty-two passages was scored for the following: unique words, mean number of letters per word, percent of words with more than five letters, mean number of words per sentence, 101 spelling variables and other variables as found in two previously developed readability formulas. Passage difficulty was calculated and used as the criterion, all other variables being correlated with it.

After examination of the results, a readability formula was developed using a weighted combination of the following: percent of unique words, mean number of words per sentence and ten spelling variables. This formula had a correlation (R) of .96 with passage difficulty and a R^2 equal to 0.92. Similar readability formulas were also developed for the sixth and ninth grade levels.

Major outcomes and interpretations

1. A readability formula was developed which predicted the percent of comprehension of a passage with a high degree of accuracy as indicated by R^2 equal to 0.92. This was very much better than the Dale-Chall or Fry formula results for the same passages.
2. This formula gives a directly useful measure of the readability level of a passage at the senior high school, ninth and sixth grade levels.

INNER LONDON EDUCATION AUTHORITY: LITERACY SURVEY

Great Britain
1969

Purpose

This literacy study was designed to (1) discover the extent of reading difficulties of all eight-year-olds in the Inner London Authority, (2) obtain a picture of both the problems seen by Head teachers and remedial action already given in junior schools, and (3) obtain a picture of London's standards as compared to those of Great Britain.

Methodology and instrumentation

The sample consisted of 31,308 eight-year-old children in the Inner London authority in October, 1968. The following procedures were used:

1. All junior schools in the authority gave a standardized National Foundation for Educational Research in England and Wales (NFER) Sentence Completion Test to all eight-year-old students in their schools. All blanks were at the end of the sentences and five choices were provided for each sentence.

2. The participating schools were asked to complete a questionnaire for each child giving details of educational and social background.

3. Each school Head teacher was asked to complete a questionnaire concerned with the amount, type and structure of remedial help given in the school as well as the Head teacher's view of the problems involved and the merits of different kinds of help in the area of reading comprehension.

Analysis

Results were given as frequency counts or percentages.

Major outcomes and interpretations

1. On the average the sample was about six months behind the national standardization group.

2. Factors found to influence achievement were immigration status, home background, family background, pre-school education, school stability and remedial help received.

3. The testing instruments used were of the traditional multiple-choice variety. This caused validity questions to be raised with respect to the findings of the study.

INNER LONDON EDUCATION AUTHORITY: LITERACY SURVEY JUNE 1971 FOLLOW-UP

Great Britain
1972

Purpose

The 1971 follow-up study was designed to provide further information on the students surveyed in 1968 and to provide a description of some of the problems likely to face the secondary schools soon to receive these students.

Methodology and instrumentation

The sample consisted of 31,731 children born between September 2, 1959 and September 1, 1960. Of these children, 26,202 had been surveyed in the 1968 survey.

The following procedures were used.

1. All children in the sample were given a parallel version of the NFER Sentence Completion Test which was used in the 1968 survey.

2. Ten percent of the sample was given either the Neale Analysis of Reading Ability or the NFER Reading Comprehension Test (DE). The Neale Analysis of Reading Ability was an individual test composed of six passages which formed a continuous reading scale for children aged six to thirteen years. The child read the passage aloud, then answered questions about it. These provided a measure of reading accuracy and a measure of reading comprehension. The Reading Comprehension Test (DE) was designed for children in top junior and first year secondary. It was chosen partly to discriminate between children at the top end of the scale on the Sentence Completion Test and to have a second group test for comparison purposes. The children were asked to read and draw inferences from eight passages taken from works of literature.

3. The Rutter behavior scale was completed for each child in the ten percent sample.

Analysis

Results were presented as frequency counts or percentages.

Major outcomes and interpretations

1. On all of these reading tests the London children were reading at a level about ten months younger than their mean age of eleven years - three months. While the mean standard score on the Sentence Completion Tests remained the same for 1968 and 1971, the gap between chronological and reading age widened in the three years.

2. The results of the behavior scale showed that a higher proportion of the children were rated "abnormal" than had been found in Rutter's general population.

INNER LONDON EDUCATIONAL AUTHORITY: LITERACY SURVEY 1973-74 FOLLOW-UP

Great Britain
1975

Purpose

The 1973-74 follow-up was designed to provide further information about the performance and progress of the students surveyed in 1968 and 1971, now thirteen to fourteen years old, and indicate some of the problems facing secondary schools.

Methodology and instrumentation

The sample consisted of 25,955 students in the ILEA who were thirteen to fourteen years old. Of these children, 20,983 had been tested in 1971. The following procedures were followed at all participating schools.

1. The students were given the Reading Comprehension Test, developed by the NFER and restandardized in 1973. This was a multiple-choice test with five items following each passage.

2. All secondary schools were asked to complete a short questionnaire giving details of the organization and teaching of English in the school and of the qualifications and experience of the teachers involved.

3. Each teacher was asked to complete a questionnaire for each child dealing with the child's reading ability and remedial help received.

4. Teachers completed the Rutter behavior scale for ten percent of the sample.

Analysis

Results were presented as frequency counts or percentages.

Major outcomes and interpretations

1. It appears that the ILEA children improved their position since 1968 and 1971. It was felt this was due to considerable resources and effort spent by the secondary schools in their attempts to deal with the problems posed by the very low reading standards of many of their pupils.

CONTEXT IN READING AND LISTENING: A COMPARISON OF CHILDREN'S ERRORS IN CLOZE TESTS

Mary H. Neville
A. K. Pugh

Great Britain
1974

Purpose

The purpose of this study was to compare children's performarce on similar cloze tests of reading and listening with regard to incidence and type of error.

Methodology and instrumentation

Two forms of the Gapadol (GAP) test of silent reading comprehension were administered. This was a cloze test in which approximately every tenth word (linguistically random) was deleted. Although of Australian origin, it had been standardized in the United Kingdom and was of known reliability and validity. Synonyms were not allowed in scoring. The same test was recorded on tape leaving a ten second gap for each missing word. This was then used as a listening test.

The sample consisted of sixty-six children from a Roman Catholic primary school in Yorkshire, England, aged between 8 years 10 months and 10 years 9 months. These children were split into two groups of thirty-two and thirty-four. The two groups were tested with first a reading test and then a parallel listening test. Errors were categorized as omissions, inappropriate, or incorrect but linguistically appropriate.

Analysis

Detailed analysis of type and frequency of errors was presented.

Major outcomes and interpretations

1. Generally, when the children listened they found the task of completing the cloze test more difficult than when they read. When a blank occurred near the beginning of a sentence they could not anticipate the word. However, when children had sufficient preceding information in both listening and reading they did try to choose words which were syntactically and to a lesser extent, semantically appropriate.

2. Scores for the listening tests for both groups were significantly lower than the reading scores.

3. There was no significant difference between the groups in their performance on the GAP reading test and a previously administered NFER reading test which was a multiple-choice sentence completion test with the blank at the end of the sentence.

4. The study substantiated the suggestion that the beginner in reading may have some difficulty in using context when he reads a new text, particularly when the unknown words come at, or near, the beginning of a sentence.

5. This study suggested there is probably a reading level below which the cloze technique may not be appropriate as a measure of reading comprehension.

CONTENT IN READING AND LISTENING: VARIATIONS IN APPROACH TO CLOZE TASKS

Mary H. Neville
A. K. Pugh

Great Britain
1976-77

Purpose

The major purpose was to compare performance on a normal cloze reading test with listening cloze test scores (Neville and Pugh 1974) as well as with performance on a reading cloze test where the message was restricted in display in a manner similar to the listening test.

Methodology and instrumentation

As in the 1974 investigation, two forms of Gapadol (GAP) test were used. Both forms were again modified to a listening test form. A further modification for the third version of the test involved printing the test in small booklets where on each page of the booklet, the words of the text appeared sequentially until a gap occurred. This version of the reading cloze test simulated the listening test in that the subject could not look ahead or back for clues to fill in the gap at the end of each page.

One hundred and sixty students aged nine to ten years from a middle school in the northern region of England wrote the regular GAP Reading Comprehension Test. (Eighty wrote form Red, and eighty wrote form Blue.) One week later, four groups totaling 130 of those students wrote either the listening test or the modified restricted display test.

Analysis

All the test papers were scored and errors were categorized. Means, standard deviations and analyses of variances were calculated for each group.

Major outcomes and interpretations

1. General analysis confirmed the findings of the 1974 investigation. The listening test scores were significantly lower than the reading test scores.

2. Restricted display test scores and listening test scores were very similar.

3. Poorer readers performed similarly in all three testing modes. Better readers did much better on the regular GAP cloze test than in the other two modes.

4. Good readers differed from poor readers in types of errors made only in the regular cloze task but not in the listening or restricted tasks.

AN OBJECTIVE PROCEDURE FOR CONSTRUCTING DOMAIN-REFERENCED TEST ON LITERAL COMPREHENSION

Robert P. O'Reilly

New York
1975

Purpose

The purpose was to develop, based on a modification of the cloze technique, a virtually objective procedure for assembly of passages and items in a multiple-choice format that used the computer to partially aid in the test construction assembly process.

Methodology and instrumentation

Several relevant domains from grade one to twelve were randomly sampled to select 1,500 seventy-word passages. All passages were converted to standard cloze format deleting every fifth word if it was a noun, verb, adjective, or adverb. Five distractors were then selected for each deletion using a compilation of core and content word lists. A computer program accessed this matter list and assigned sets of distractors to deletions in a passage on the basis of the type of word deleted. The type was determined by the word's form class, grade level at which the word first frequently appeared, and whether the word was a general or core word. The distractors were systematically scrambled on the output, with the position shown for assignment of the correct answer. Passages were then scaled by grade level by systematically eliminating passages which deviated greatly from the central tendency of a grade. The result was an overlapping but graduated increase in passage difficulty from grade to grade.

Although three item formats were developed, over ninety percent of the material was in the format of a seventy-word passage with ten deletions, and five distractors per deletion. At least fifty non-overlapping fifty-item test forms could be assembled from the reading test passages alone. Texts in language arts, social studies, science, mathematics, reference materials, citizen materials, and materials read by consumers have been converted in this manner.

Major outcomes and interpretations

1. These testing materials are currently being validated on over 5,000 students.

2. It is expected that subsequent calculation of all passages on a common scale should enable any user to assemble a tailored test form to assess the development of literal comprehension for a given student.

3. Any test form assembled from the pool will then be comparable to any other test form that might be assembled using the same specifications.

THE POST ORAL READING CLOZE TEST: NEW LINK BETWEEN ORAL READING AND COMPREHENSION

William D. Page Illinois
1975

Purpose

This study investigated the relationship between conventional cloze test performance with no prior contact with the passage and a post oral reading cloze test with test administered immediately following an oral reading of the passage by the subject.

Methodology and instrumentation

Five matched pairs of passages each with 250 to 303 words at each of five levels of difficulty were selected from a previous Bormuth study (1973-74). Cloze tests of fifty or more items were prepared in which a fifteen-space line was substituted for every fifth word. Subjects were students from grades two to six in one Chicago school. Two males and two females from high, middle, and low ability were selected at each grade level yielding sixty subjects in total. Each subject read a passage orally, responded to a post-oral reading cloze (POC) test and then to a conventional cloze (CC) test on a matched passage. This procedure was repeated with each subject for each of five passage difficulty levels. The administrations were from easiest to most difficult. Order of administration was balanced for the CC test, POC test, sex, grade and ability.

Analysis

Responses corresponding to the deleted word were scored as correct. Misspellings were scored correct unless the response could be construed to be another word. Raw cloze scores were transformed to percentages of the total items per passage. Correlations, ratios and regression equations were calculated for the total groups, each grade, sex, and ability groups. They also were calculated for scores grouped at each passage difficulty level.

Major outcomes and applications

1. CC test scores can be predicted reasonably well from POC test scores.

2. The relationship appears to be strong enough to warrant use of POC tests as comprehension criteria in oral reading research.

3. POC tests alone are weak diagnostic tools for individuals but are useful as a source of corroborative information in a cluster of other measures.

4. Because CC tests predict literal comprehension as measured by multiple-choice comprehension tests (Bormuth, 1969) the POC tests represent a valuable link between oral reading performance and comprehension performance.

CONTEXT-DEPENDENCE OF ITEMS DESIGNED TO MEASURE THE ABILITY TO DERIVE THE MEANING OF WORDS FROM THEIR CONTEXT

Fred Pyrczak

California
1976

Purpose

The purpose of this study was to estimate empirically the context-dependence of items that were drawn from a number of published tests. These questions were designed to measure the ability to derive the meaning of words from context.

Methodology and instrumentation

In the first step, thirty-two items in seven different reading tests were classified as vocabulary-in-context items. These items were assembled into two forms and presented without the associated selections. Examinees were 159 students from six senior high school classes who wrote one of the forms of the test.

Analysis

Means and standard deviations of the two forms, as well as of each item, were calculated and compared to a completely context-dependent mean of 4.35. An index of context-dependence for each item also was calculated.

Major outcomes and interpretations

1. Except for items in the Davis Reading Test, vocabulary-in-context items in the published reading comprehension tests examined lacked context-dependence. This meant that students may answer correctly many items without first reading the passage.

2. Since the purpose of these specific items was to measure vocabulary-in-context, the presence of mostly non-passage-dependent items creates doubt as to the validity of many of the publisher tests.

THE CLOZE PROCEDURE: A NEW TOOL FOR ADULT EDUCATION

Robert D. Robinson Missouri
1973

Brief refiew of the literature

This article had an excellent reference list for developing cloze questions and using the cloze tehcnique. Some of the major events and dates listed where these.

First published paper on cloze procedure: 1953

Research on methodology: 1953, 1956, 1965, 1967

Early studies which suggested the cloze procedure as a viable alternative to comprehension testing; cloze correlated to standardized tests: 1959, 1963, 1968

Tables developed which allowed comparisons of difficulty levels of cloze and multiple-choice passages: 1967, 1968, 1969

Use of cloze as a substitute for readability formulas in the measurement of passage difficulty: 1964, 1967, 1968

Purpose

This study was designed to evaluate the cloze technique on an adult population by testing three types of cloze construction as predictors of a standardized reading test, evaluating cloze procedure in relation to several readability formulas and soliciting opinions of the cloze procedure from examinees.

Methodology and instrumentation

Fifty-seven inmates of a United States federal penitentiary were used in this study. The subjects, each having a mean I.Q. greater than or equal to ninety and reading grade scores between 4.0 and 7.9, wrote three test types.

1. Cloze Deletion Test

   A test was developed and administered to investigate three types of cloze deletion: multiple-choice deletion, 1-5 deletion and noun-verb deletion. There were fifty deletions of each of the three types used. Scoring was based on exact replication of the original test, except for spelling.

2. Cloze Ranking Tests

   A 1-5 cloze test consisting of fifty deletions was constructed on each of seven adult passages and administered to subjects. These same selections were ranked for difficulty using readability formulas previously developed. The purpose of these tests was to determine the ability of the cloze procedure to rank a series of selected passages in comparison to the rankings of several readability formulas.

3. Cloze Opinion Survey

   Ten questions using the five response mode "strongly agree" to "strongly disagree" were constructed and administered to determine opinions toward the cloze items.

Analysis

1. Stepwise regression analyses were computed using the predictor variables of intelligence, cloze multiple-choice, cloze 1-5 deletion, and cloze noun-verb deletion to predict various subtests of the Stanford Achievement Tests (SAT).

2. A rank order coefficient of correlation was used to determine the degree of association between rankings of the seven prose passages.

Major outcomes and interpretations

1. The 1-5 deletion pattern seemed to predict scores on the paragraph meaning and word meaning subtests of the SAT as well as group test scores of verbal ability.

2. None of the rank order correlations were significant at either the 0.10 or 0.01 levels.

3. The results of the cloze opinion survey indicated the subjects had a positive attitude toward reading, wanted to improve their reading skills and found the cloze testing to be understandable, interesting and a source of new information.

4. The study suggested that a series of short cloze passages could be successfully used as a placement instrument for adults.

PASSAGE-DEPENDENCY DATA IN THE SELECTION OF READING COMPREHENSION TEST ITEMS

Dr. Henry Scherich
Dr. Gerald Hanna

California
1976

Purpose

The major purpose was to study the passage-dependency of a newly developed reading test, the Nelson Reading Skills Test (NRST), developed by Hanna and the Houghton-Mifflin Company.

Methodology and instrumentation

Nine test booklets containing 327 items were administered in October, 1975 to 1,600 grade four and grade six students from five widely separated school districts in the United States.

The first subtest of each booklet contained a normal reading comprehension test with reading passage and questions about the passage; the second subtest contained approximately an equal number of questions on another topic but without reading passages. Every question appeared twice - once with its reading passage and once in another booklet, without its reading passage. Each subtest had a twenty-five minute time limit; all items were multiple-choice with four options. Both reading comprehension and words-in-context items were used in the study.

Analysis

Statistics were calculated to determine percentage of students selecting the correct answer with the passage-in (PI), the percentage selecting the correct answer with the passage-out (PO), and a passage-dependency index.

Major outcomes and interpretations

1. Generally, the results of the passage-dependency study supported the validity of the NRST items.

2. The passage-dependency results helped in the selection and/or changing of items in ways believed to be beneficial in constructing a test with greater validity.

DETERMINING THE PASSAGE-DEPENDENCY OF COMPREHENSION QUESTIONS IN FIVE MAJOR TESTS

J. Jaap Tuinman

Indiana
1974

Purpose

Major purposes of this study were to (1) investigate the degree that reading comprehension test items were passage-dependent, (2) show that passage-dependency was related to the educational growth of the respondents, and (3) produce reliable statistics on the passage-dependency of items used in the study.

Methodology and Instrumentation

Five major reading tests were administered to 1,200 students who were not allowed to read the related passage. Control data was obtained by administering each test in its regular format to a sample of 600 students. Students for both the experimental and control conditions were selected from a number of grades in ten school systems in Indiana.

Analysis

Means, standard deviations, correlations (KR_{20}), and standard errors of measurement (SEM) were calculated for each test in both passage-in and passage-out modes. The same analyses were also done separately by grade for each condition.

Major outcomes and interpretations

1. Commercially marketed tests were found to vary considerably in the degree to which their items were passage-dependent.
2. None of the five tests tested even approached passage-dependency.
3. Tests designed for grades four to six showed a consistent decrease in passage-dependency from fourth to sixth grade students.
4. An index for determining passage-dependency was recommended.

SIX-SUBJECT STUDY FOR THE INTERNATIONAL ASSOCIATION FOR THE EVALUATION OF EDUCATIONAL ACHIEVEMENT (IEA)

Sweden
1966-1976

Purpose

In 1965 the International Association for the Evaluation of Educational Achievement (IEA) inaugurated a cross-national survey of achievement in six subjects: Reading Comprehension, Literature, Science, English as a foreign language, French as a foreign language, and Civic Education.

The overall aim of the project was to use international tests in order to relate student achievement and attitudes to instructional, social and economic factors and, from the results, to establish generalization of value to policymakers worldwide (Schwille and Marshall, 1976).

Methodology and instrumentation

From eight to nineteen countries participated in each survey by testing students at one or more levels. These levels were ten years, fourteen years, and pre-university.

The Six-Subject Survey was broken into four stages.

1. 1966-70 - Construction and pretesting of instruments,
2. 1970-73 - Field work and analysis for Science, Reading Comprehension, and Literature,
3. 1971-74 - Field work and analysis for Civic Education, English as a foreign language and French as a foreign language,
4. 1973-75 - Rebuilding of files and preparation of documentation for archiving and distribution (Schwille and Marshall, 1976).

A brief description of the Literature Tests (Husen, 1974), Reading Comprehension Tests and Reading Speed Test (Husen, 1974) are included in this summary.

Literature Tests

Each of the five tests consisted of a short story and a series of thirty-eight or thirty-nine multiple-choice questions. The first twenty items were questions which might be asked about the story. Students were to choose the five questions which they thought were the most important questions to ask about the story. The remaining

items asked questions about the short story. Some items were marked as more important than others and students were asked to pay particular attention to those items. The following stem is an example of a more important item.

> Lines 74 to 77 may be said to be one of the points at which the direction of the story turns. Which of the following best summarizes the turn?
>
> (Husen, 1974, p. 36)

Reading Comprehension Tests

Each test was made up of four short stories. Each story was followed by five to seven multiple-choice questions. Most questions were comprehension category questions.

Reading Speed Test

This test was composed of a story divided into forty parts. At the end of each part there were three words. Students were asked to underline the word that fit into the story. Students completed as many of the parts as they could in four minutes. An example of one part is:

> Spot saw Peter throw a stone. He started to run after the stone. He did not see the stream. He fell into the hold mud water.
>
> (Husen, 1974, p. 31)

Major outcomes and interpretations

1. Twenty-one IEA National Centers collected student data on achievement, home and socioeconomic backgrounds, interests and attitudes as well as data on teachers and schools. Approximately 250,000 students, 50,000 teachers and 10,000 schools were involved.

2. Major conclusions were presented in Thorndike's (1973) report on this project. Those conclusions dealing with the reading survey were:

   (a) With complete consistency the three developing countries (Chile, India, Iran) fell far below those that had a relatively high level of economic development and a long-standing tradition of universal education.

   (b) In the developed countries an appreciable prediction of reading achievement of individual students--and an even more substantial prediction of the average reading achievement of children in school--was provided by information about their home and family backgrounds.

(c) These two findings were basically consistent with one another because the developed countries differ from the developing ones perhaps most sharply in just those characteristics that characterize the children in the developed countries who read well.

(d) The study provided very little evidence of the impact of the school or of specific factors on the progress of students in reading.

3. Some of these conclusions have been questioned as to their accuracy since Thorndike presented them (Downing and Dalrymple-Alford, 1974-75).

APPENDIX B:

SUMMARIES OF RESEARCH RELATED TO ASSESSMENT OF WRITING

WRITING: NATIONAL RESULTS

National Assessment of Educational Progress — Colorado 1970

Introduction

Exploratory Committee on Assessing the Progress of Education (ECAPE) established in 1964 by the Carnegie Corporation, devised a plan calling for the periodic assessment of educational outcomes in ten subject areas at four age levels: nine-year-olds, thirteen-year-olds, seventeen-year-olds, and young adults between twenty-six and thirty-five years of age. Writing was one of the areas to be assessed. Educational Testing Service of Princeton, New Jersey, under contract to National Assessment of Educational Progress (NAEP), developed objectives and exercises in writing. They worked with subject matter experts, educators, and lay people concerned with educational planning. Four broad objectives of writing were defined. They were:

1. writing to communicate adequately in a social situation;
2. writing to communicate adequately in a business or vocational situation;
3. writing to communicate adequately in a scholastic situation; and
4. appreciating the value of writing.

Each of these objectives was elaborated in more specific terms and examples of types of behaviors expected at each of the four age levels were stated. All objectives and exercises were reviewed and accepted by ECAPE before data collection began.

Gathering the data

Data was collected from approximately 24,000 nine-year-olds, 28,000 thirteen-year-olds, 28,000 seventeen-year-olds, and 8,600 young adults. Respondents were selected according to a nation-wide probability sampling plan. Results were then weighted to correct for sample biases. Exercises were divided into ten, thirteen, fourteen, and ten booklets for the four ages respectively. Approximately 2,000 responses were obtained for each exercise at each of the ages nine, thirteen, and seventeen and from 650 to 900 responses for the young adult range.

Most exercises were group-administered with instructions and exercises presented using a tape recorder as students followed along in their packages. Individually-administered exercises were given by a trained interviewer who read each exercise aloud and then

recorded the response. All adult booklet packages were individually-administered.

Types of exercises, scoring, and reporting

Three types of writing exercises. Type one called for a short answer or essay. Answers were coded into predetermined categories according to a very specific scoring guide. If all information needed to communicate the intent of the message was included, the response was scored "acceptable."

Type two asked the respondent to indicate his answer by marking the appropriate space. All exercises of this type were written for Objective IV, Appreciate the Value of Writing. The percentages of respondents saying "yes" and "no" to each subsection or combination of subsections were reported.

Type three asked the respondent to write about a given topic. The responses were assessed for "writing ability" as a whole. Adequate grammar, appropriate word choice, originality, ability to express ideas and depth of thought all contributed to this assessment. Results were reported by showing actual examples of essays demonstrating writing abilities reached by approximately 15%, 50%, and 85% of the respondents.

In each of the three types assessed, a number of exercises were administered at more than one age. These overlapping exercises sometimes differed slightly in wording but involved essentially the same task. The percentage of respondents able to perform a given exercise acceptably was compared between ages where applicable.

Results for each age level. The percentage of respondents who were able to complete each writing task successfully was shown by age level and by objective within each age. The results for the 1969 administration for the thirteen-year-old age sample are illustrated in Figure 9. Similar charts were prepared for the three other samples. Samples of exercises and their accompanying responses, marked in three categories, for two of the samples are given in Figures 10 to 20.

Samples

Percent Acceptable	OBJECTIVE			
	I Social	II Business	III Scholastic	IV Appreciation
95-99				
90-94	●			
85-89				●●●○
80-84				
75-79	●	○		○
70-74				
65-69	●			
60-64				●○
55-59				○
50-54	○			
45-49	○	●		
40-44				
35-39				
30-34				
25-29		●●		
20-24				
15-19	○			●○
10-14		○		
5-9				
0-4				
Essays	●		●●○○	

● Released

○ Not released

Each circle represents one exercise

(Hazlet, 1970, p. 23)

Figure 9
Number of NAEP Writing Exercises at Different Levels of Success for Age 13 Sample.

EXERCISE 203

(13:8-17) (Similar to Exercises 103 and 301)

You are going to hear a telephone conversation between two boys, Al and Ben. During the conversation, you will discover that Ben is going to have to write a note to his mother. Listen carefully to find out the things that Ben will have to say in his note. (Children then listened to the conversation reproduced below. It was not printed in the children's booklet.)

AL: Hey, Ben, this is Al. I called you to remind you you're supposed to come down to my house for supper tonight before the game.

BEN: Oh yeah, I remember. Your mother and father are going to pick me up on the way back from the barber shop. What time do you think you'll get here?

AL: Oh, around 5 o'clock.

BEN: O.K. I'd better leave a note for my mother. She won't be home until 5:30 and maybe she's forgotten that I won't be eating at home.

AL: You'd better remind her of the ball game, too. She's supposed to pick us up afterwards, isn't she?

BEN: That's right. Your father is taking us to the game but my mother is bringing us home. Thanks for reminding me. I'll write the note right away.

You will now have another chance to hear the conversation between Al and Ben. Listen to it carefully again and then, when the boys have finished talking, write the note that you think Ben should leave for his mother.

(Approximately 1/2 page of lined space was provided for the response.)

(Hazlett, 1970, p. 24)

Figure 10
An "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective I Social, Type One.

Acceptable: The response must:

1. be in the form of a note that states
   a. he is going to Al's house for dinner, and
   b. Al's father is taking them to the ball game (or the reader must be able to infer this information), and
   c. his (Ben's) mother should pick them up after the game.

Not acceptable: Omission or error in any of the above pieces of information, or copying the instructions.

(Hazlett, 1970, p. 24)

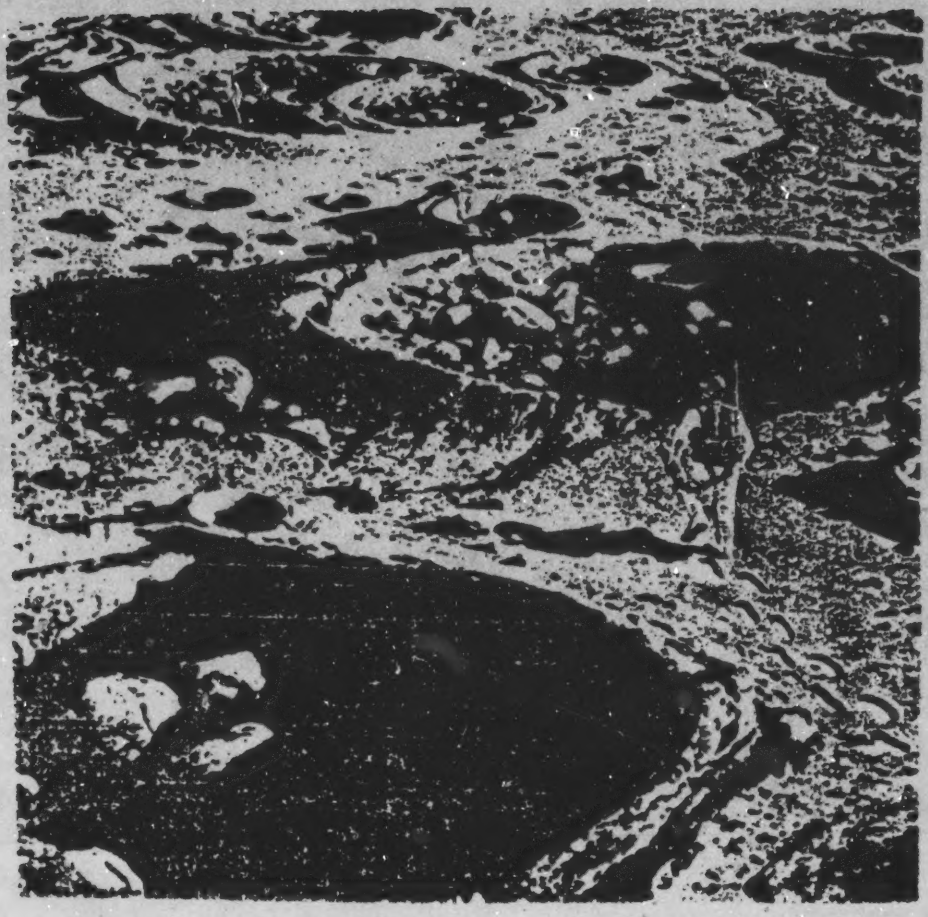
Figure 11
Scoring of "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective I Social, Type One.

Age 13

67%	Acceptable
32	Not acceptable
1	No response
	Percentage providing each piece of information:
84%	mention that he is going to Al's house for dinner
81	mention or infer that Al's father is taking them to the ball game
84	mention that his (Ben's) mother should pick them up after the game.

(Hazlett, 1970, p. 25)

Figure 12
Results of "Easy" Exercise Prepared for Thirteen-Year-Olds on Objective I Social, Type One, for the United States Sample.



#E-3. (Age 9: 4-19)

Here is a picture of an astronaut on the moon. Look at the picture for a while and think about what is happening. Now, pretend that YOU were the astronaut, and write a story about your walk on the moon.

(The rest of the page provided lines for the responses. Students could use the back of the page if they wished.)

(Hazlett, 1970, p. 61)

Figure 13
An Essay Prepared for Nine-Year-Olds on Objective III Scholastic, Type Three.

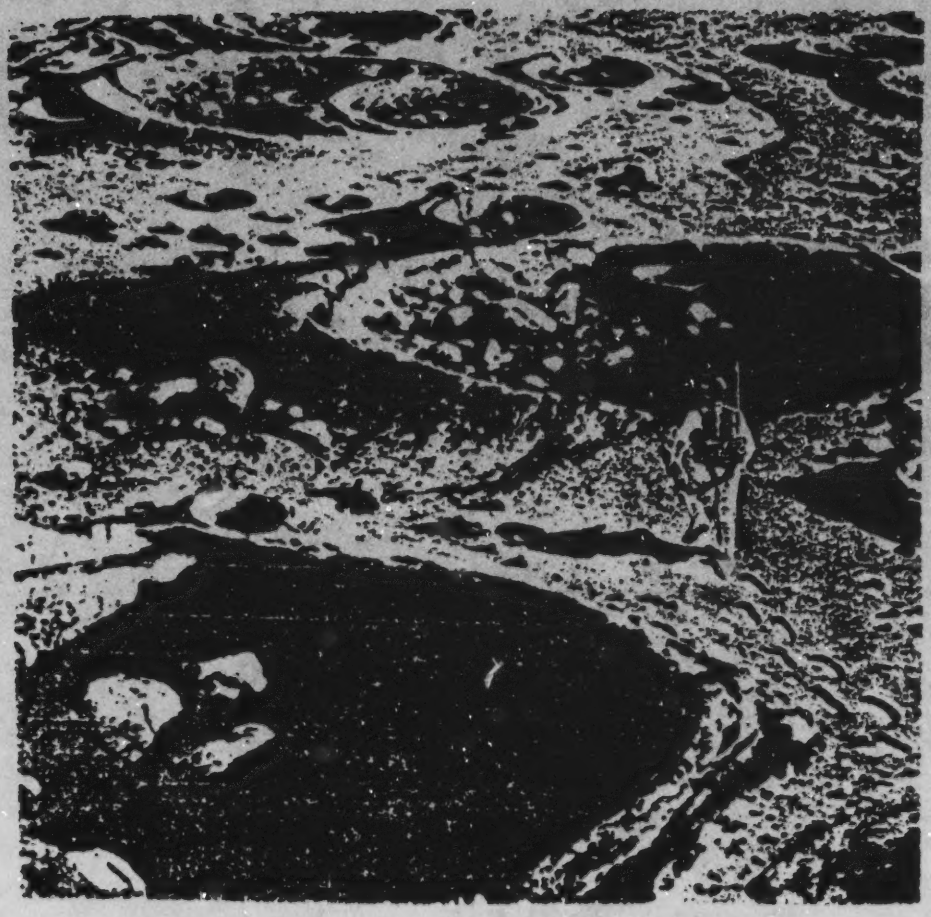
When I was walking on the moon I was scared a little because it was so dark on the moon. I walked inbetween two craters they were very big, big enough to hold a little house. There was a shadow of a giant crator, and boulders big enough to crash right through a steel wall. So its time for me to go back to the L.E.M. now
Good-By

The ship landed with a crash and my partener stayed in the ship incase any thing hupend I got out of the ship but I was a little afraid because man has never walked on the moon when I got out the sea I felt quite funny on my feet and there I was the first man on the moon I walked around for a while and stick the American Flag on the moon and got into the ship and we took off on our way home the Earth and dad I have a story to tell the head of the space program when we got home.

When I got out of our spaceship I explored the planet. When I was walking I saw some light on the other side of the planet, it looked like Uranium. I got into my spaceship and blasted off. Then I landed it where the light was comming from. When I got out the light I saw was in a cave I went into the cave. And went out the other side. when I saw the light it was comming out of a volcano. Then at that second something strange happened the volcano erupted. I ran back to my spaceship. When I got back to my spaceship the lava was behind me. Then I blasted off.

(Hazlett, 1970, p. 62)

Figure 14
Sample Responses of the Better Papers Written by Nine-Year-Olds on the Previous Exercise. Eight Percent of the Nine-Year-Olds Wrote Papers at Least as Good as These.



#E-3. (Age 9: 4-19)

Here is a picture of an astronaut on the moon. Look at the picture for a while and think about what is happening. Now, pretend that YOU were the astronaut, and write a story about your walk on the moon.

(The rest of the page provided lines for the responses. Students could use the back of the page if they wished.)

(Hazlett, 1970, p. 61)

Figure 13
An Essay Prepared for Nine-Year-Olds on Objective III Scholastic, Type Three.

When I was walking on the moon I was scared a little because it was so dark on the moon. I walked in between two craters they were very big, big enough to hold a little house. There was a shadow of a giant crator, and boulders big enough to crash right through a steel wall. So it is time for me to go back to the L E M. now Good-by

The ship landed with a crash and my partener stayed in the ship incase any thing hupend I got out of the ship but I was a little afraid because man has never walked on the moon when I got out the seal felt quite funy on my feet and there I was the first man on the moon I walked around for a while and stick the American Flag on the moon and got into the ship and we took off on our way home the Earth and did I have a story to tell the head of the space program when we got home.

When I got out of our spaceship I explored the planet. When I was walking I saw some light on the other side of the planet, it looked like Uranium. I got into my spaceship and blasted off. Then I landed it where the light was comming from. When I got out the light I saw was in a cave I went into the cave. And went out the other side when I saw the light it was comming out of a volcanoe Then at that second something strange happened the volcanoe erupted. I ran back to my spaceship. When I got back to my spaceship the lava was behind me then I blasted off.

(Hazlett, 1970, p. 62)

Figure 14
Sample Responses of the Better Papers Written by Nine-Year-Olds on the Previous Exercise. Eight Percent of the Nine-Year-Olds Wrote Papers at Least as Good as These.

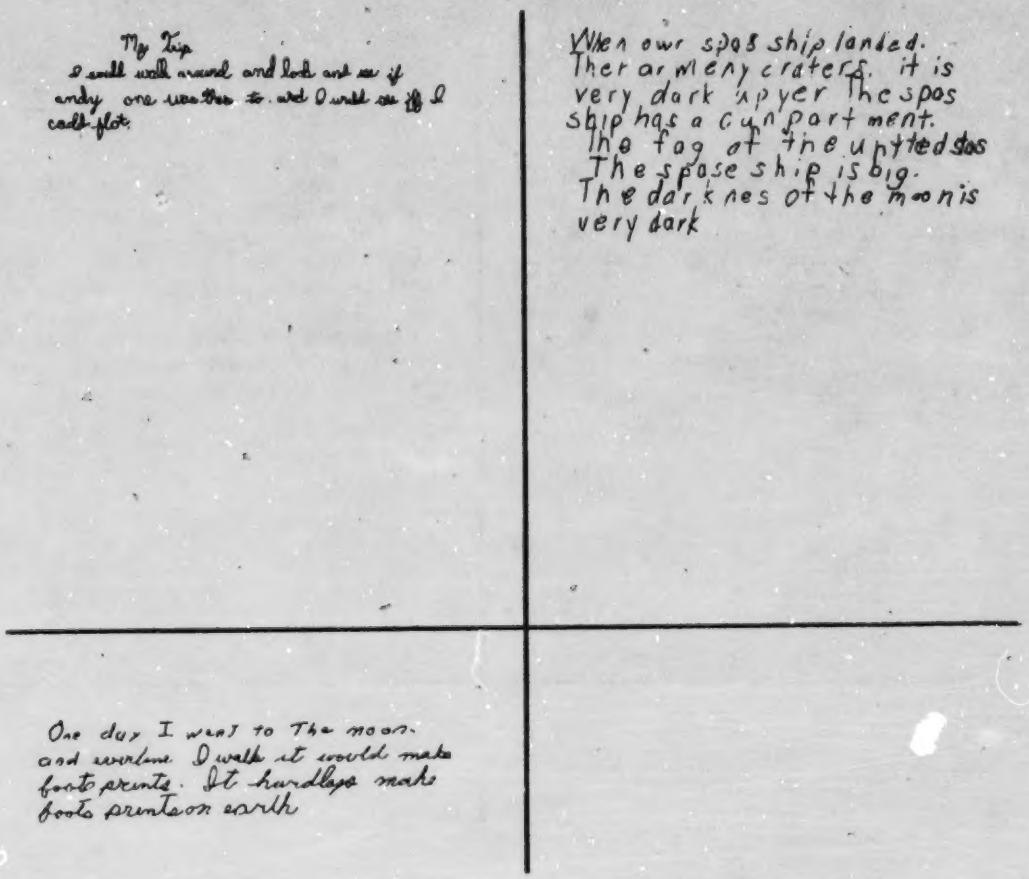
My visit to the M...
I am going near to the moon. I may land soon. I have landed on the moon I see a big crater. It has a lot of spooky rocks in it. I am going to get some. I am going back to the ship and leave I am going back to earth. Good-by moon.

We are walking on the moon and we find some moon rocks. Then the other guy almost stepped in a gravle. Then we walked up farther and saw some strange things we saw a big rock. It was all different colors. We walked up farr and saw something moving it was a moon man. We talked kind of strange. Then we got on the rocket ship and herd a sound

Well my story begins on the moon. I was walking on the moon when I fell down I hurt my shelf but I got right up agin. So I went right on by and by I had to go back down to the earth.

(Hazlett, 1970, p. 63)

Figure 15
Sample Responses of the Typical Papers Written by Nine-Year-Olds on the Previous Exercise. Fifty-Nine Percent Wrote Papers at Least as Good as These.


My Trip
I will walk around and look and see if
enby one was there to. and I will see if I
can't find.
When our spos ship landed.
Ther ar meny craters. it is
very dark upyer The spos
ship has a cun part ment.
The fog of the uptted sos
The spose ship is big.
The darknes of the moon is
very dark
One day I went to The moon.
and wenlme I walk it would make
foots prints. It hardlys make
foots prints on earth

(Hazlett, 1970, p. 64)

Figure 16
Sample Responses of the Weakest Papers Written by Nine-Year-Olds on the Previous Exercise. Ninety-Seven Percent Wrote Papers at Least as Good as These.

Age 17

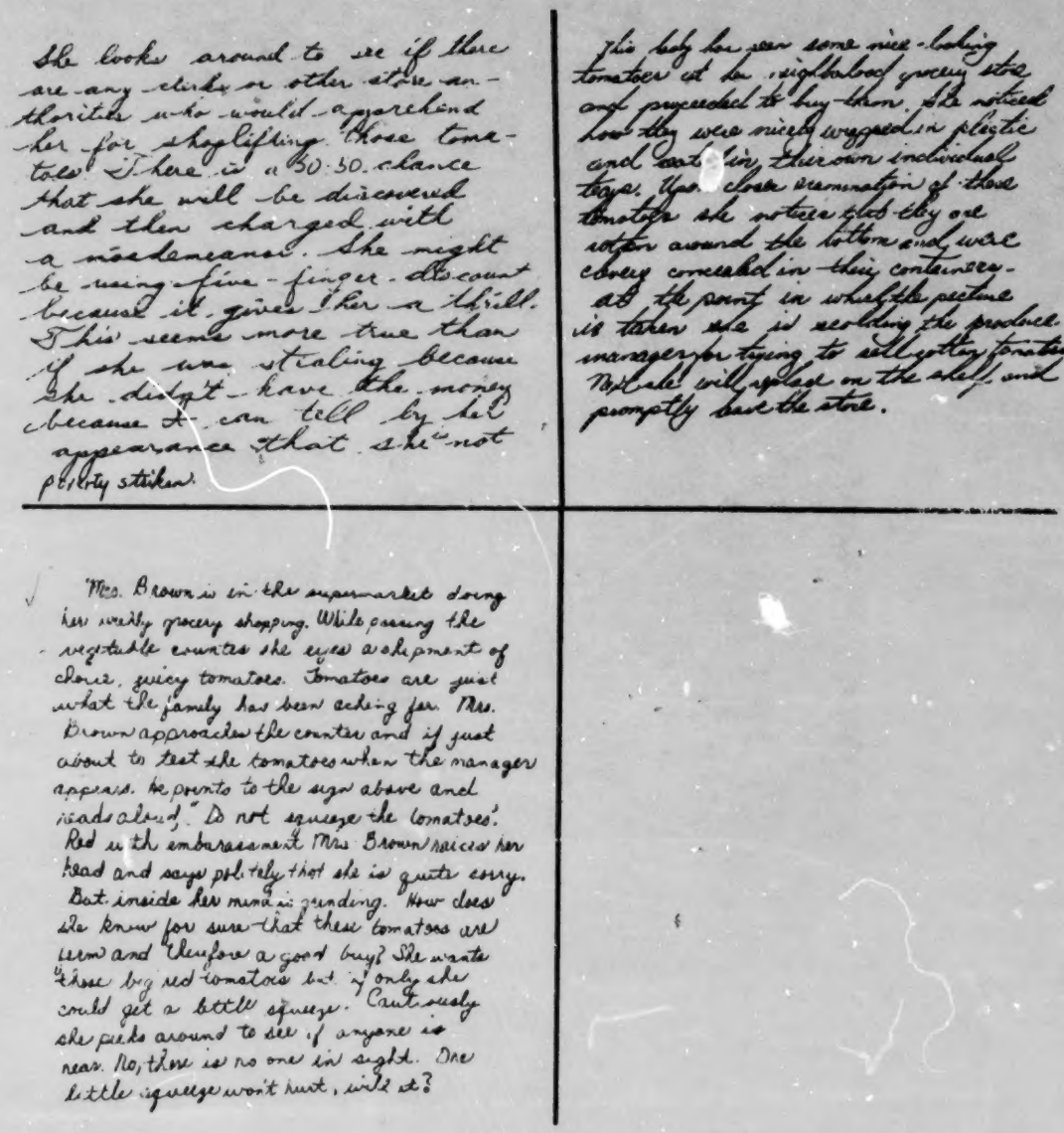
The one non-overlapping exercise released at this age showed a picture of an elderly lady holding a package of tomatoes (the picture shown students was in color).

#E-6. (Age 17: 1-11)

Here is a picture of a woman with some tomatoes. Look at the picture for a while and decide what is going on. When you have decided, write a story that tells what is happening in the picture and what is likely to happen next.
(The rest of the page provided lines for their responses. Students could use the back of the page if they wished.)

(Hazlet, 1970, p. 73)

Figure 17
An Essay Question Prepared for Seventeen-Year-Olds on Objective IV Scholastic, Type Three.



She looks around to see if there are any clerks or other store authorities who would apprehend her for shoplifting. Those tomatoes! There is a 50-50 chance that she will be discovered and then charged with a misdemeanor. She might be using five-finger discount because it gives her a thrill. This seems more true than if she was stealing because she didn't have the money because I can tell by her appearance that she's not plenty stinko.

This lady has seen some nice-looking tomatoes at the neighborhood grocery store and proceeded to buy them. She noticed how they were nicely wrapped in plastic and each in their own individual bags. Upon close examination of these tomatoes she noticed that by one rotten around the bottom and was cleverly concealed in their containers. At the point in which the picture is taken she is scolding the produce manager for trying to sell rotten tomatoes. Mr. Kale will upload on the shelf, and promptly leave the store.

Mrs. Brown is in the supermarket doing her weekly grocery shopping. While passing the vegetable counter she eyes a shipment of choice, juicy tomatoes. Tomatoes are just what the family has been asking for. Mrs. Brown approaches the counter and is just about to test the tomatoes when the manager appears. He points to the sign above and reads aloud, "Do not squeeze the tomatoes." Red with embarrassment Mrs. Brown raises her head and says politely that she is quite sorry. But inside her mind is grinding. How does she know for sure that these tomatoes are ripe and therefore a good buy? She wants those big red tomatoes but if only she could get a little squeeze. Cautiously she looks around to see if anyone is near. No, there is no one in sight. One little squeeze won't hurt, will it?

(Hazlett, 1970, p. 74)

Figure 18
Sample Responses of the Better Papers Written by Seventeen-Year-Olds on the Previous Exercise. Fourteen Percent Wrote Papers at Least as Good as These.

Abe looks directly about to steal them and looking around to see if any body is watching. This woman is not poor but she forgot her money and check book. She thought there were so many tomatoes that the store wouldn't miss three and she also thought that the store would be happy to let her have them because she is old and helpless. The next thing that happens is a young man sees Ma put them in her purse and tells her he saw her and if she would put them back he wouldn't tell on her and also if she promised never to steal again.

The tomatoes have just come in season and there are millions of tomatoes all over. In fact that is all she sees a tomato shop. She amazed at all the gorgious tomatoes and at the unbelievable price that they are selling them for. All of a sudden she gets an idea. She going to squeeze a tomato because she can't believe that they are real. She looks around to make sure that nobody is looking and she takes a little squeal "They're REAL!" But somebody has caught her. Little Johnie Smith is standing right there watching her. Just when she had caught Johnie had caught her, he speaks up and asked her why she was squeezing them. She tries to explain to him about how this is just about the only time of year you could ever expect to get a ripe one. She is satisfied. She bought herself a basket of tomatoes and went home to eat tomatoes, can tomatoes and she even brought tomatoe seeds to plant tomatoes. I think she likes tomatoes!

The Lady seems to be asking about her tomatoes. Like most old ladies she going to proably drive the store crazy with her questions. Such as are the fresh, are the old, or bruised. She'll proably buy the best out of the guy.

It also seems as if she is trying to find something wrong with them so she could get a discount on them.

(Hazlett, 1970, p. 75)

Figure 19
Sample Responses of the Typical Papers Written by Seventeen-Year-Olds on the Previous Exercise. Fifty-Two Percent Wrote Papers at Least as Good as These.

The woman appears to be looking to see if any one is watching her pick up the tomatoes She appears to be wanting to hide them so if she wants to steal them. It could almost be said that she has noticed someone watching her and is on the alert.

This picture is showing an old woman who is testing to see if the tomatoes are ripe or if they are soft or like the clerk is helping someone else.

While checks to see if the clerk sees her, she is still picking the tomatoes and she smells the tomatoes. She puts them back and walks at of the store

The woman is squeezing the tomatoes to make sure they are fresh, without the store manager seeing her. What will likely happen next is she will either take these tomatoes or put them back, and pick out another bunch to squeeze

(Hazlett, 1970, p. 76)

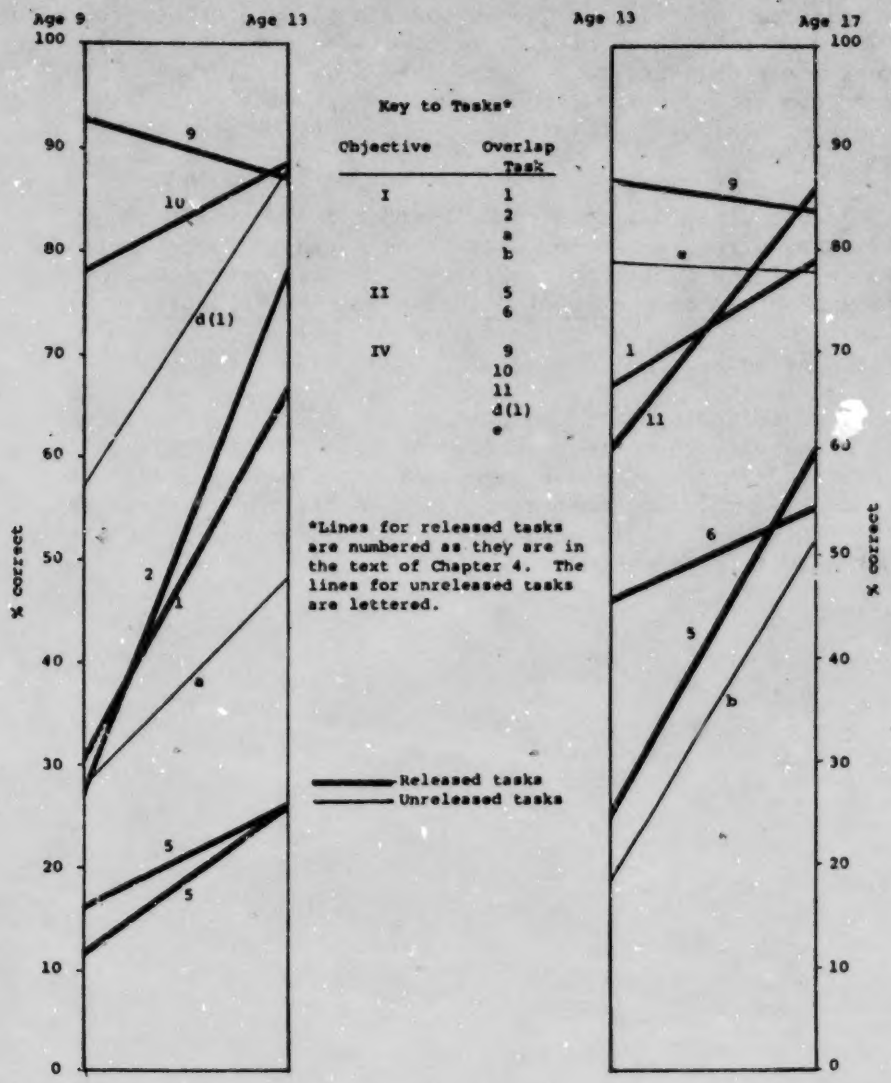
Figure 20
Sample Responses of the Weakest Papers Written by Seventeen-Year-Olds on the Previous Exercise. Eighty-Two Percent Wrote Papers at Least as Good as These.

Summary of holistic scoring procedures for essay exercises. Trained readers (1) read exercise responses which were either essays or letters, (2) formed an overall impression of "quality," (3) compared this "quality" with a set of standard responses to which scores had been assigned, and (4) assigned a corresponding score to the essay or letter. To do this, readers first read a sample of essays or letters for each exercise, ordered them from "best" to "worst," and scored them from 0 to 8. Two readers independently read and scored each essay, making the possible range of total scores per essay from 0 to 16.

Once this was completed, further analyses were done. These analyses resulted in many details being noted for each writing level. A summary of the additional analyses for these exercises follows in the summary of Writing - National Results--Writing Mechanics, 1972.

Overlapping tasks

Overlapping tasks as previously described, provide the basis for comparing performance at different ages. Almost without exception, the thirteen-year-olds did better than the nine-year-olds; the seventeen-year-olds performed better than the thirteen-year-olds. Further analysis revealed that the young adults performed more poorly than the seventeen-year-olds.



(Hazlett, 1970, p. 83)

Figure 21
Percentages of Different Age Samples Giving Acceptable Responses to Overlapping Tasks.

WRITING: NATIONAL RESULTS--WRITING MECHANICS

National Assessment of Educational Progress — Colorado 1972

Purpose

The purpose of this project was to report the results of National Assessment of Educational Progress (NAEP) writing assessment with respect to three of the writing experiences. Reports were written on those exercises requiring the writing of essays or letters.

Methodology and instrumentation

Students at age nine were shown a colored picture of a forest fire and were asked to write a story about what was happening. Time limit was fifteen minutes. No length was specified.

Students at the age of thirteen and seventeen, in groups of about twelve, were asked to write a 200-250 word essay about someone they admired. Time limit was thirty minutes.

Analysis

Four types of analyses were carried out. They are discussed below.

Grading for overall quality (holistic scoring).

1. A scale of overall quality was established by reading a sampling of essays prior to scoring. The scale used numbers from one (poorest quality) to eight (highest quality). This scale was established by three supervisory people who independently read the sample of papers and evaluated them. Papers receiving the same score from all three persons were called range finders and were subsequently used for training readers.

2. Quality was never defined by a specific set of criteria; assigned scores reflected quality relative to other papers in the sample.

3. Each paper was read and scored by two readers independently. The sum of the scores was taken as the score out of sixteen for each paper.

4. Scoring was periodically checked and discrepancies of over one point were resolved by supervisory personnel.

Counting of mechanical errors for each essay. Experienced English teachers read each essay and counted errors in punctuation, spelling, capitalization, agreement (subject/verb/gender/number, etc.), improperly used sentence fragments, run-on sentences, awkward constructions, faulty paragraphs, and word choice. A complete description of each error type was made and is available in NAEP manuals. Each essay was scored independently by two teachers and the average score for each error type was used. The error counts analysis was used to find areas where children and young adults encountered difficulties and areas where problems were minimal.

Characterization analysis. This analysis involved three groups of essays: lower quality (fifteenth percentile range), middle quality (fiftieth percentile range), higher quality (eightieth percentile range). Thus, neither the lowest nor highest essays were included. Generally, every fourth paper in the set of essays at each quality level was examined.

Judges with extensive backgrounds in the teaching of writing, literature, and linguistics issued descriptions of the essays according to the nine error categories. They described the kinds of punctuation marks used and the contexts in which they were used; they noted errors and speculated about why the errors might have appeared, described correct usages and suggested reasons why those usages appeared and others did not.

Computer analysis. Each of the essays at the three levels of quality was keypunched on computer cards precisely as the respondent had written it preserving all errors. The cards were then examined using the IBM 360/91 at Princeton University and an assembly language program called Son of MOMSR. Six specific qualities of writing were examined. They were fluency, spelling, diction, sentence structure, punctuation, and paragraph development. A list of all errors handled by the computer is given in Figure 22.

1. Number of different words in the essay
2. Total number of words in the essay
3. Mean word length in letters
4. Standard deviation of word length in letters
5. Total number of sentences
6. Number of declarative sentences
7. Number of interrogative sentences
8. Number of imperative sentences
9. Mean sentence length in words
10. Standard deviation of sentence length in words
11. Number of paragraphs
12. Mean paragraph length in sentences
13. Standard deviation of paragraph length in sentences
14. Relative pronouns
15. Head position gerunds
16. Head position adverbs
17. Head position past participles
18. Number of colons
19. Number of semicolons
20. Number of parentheses
21. Number of quotation marks
22. Obvious misspellings of common words[1]
23. Obvious misspellings of proper nouns
24. Misspellings of uncommon words[2]

[1]A common word is any word appearing at least once every two papers.
[2]Uncommon words appeared less than once every two papers.

(Slotnick, 1972, p. 157)

Figure 22
Attributes of Each Paper Examined by the Computer.

Basic assumptions underlying the measurements. The overall quality scores were used as a basis for all further analysis. All the error count measurements and characterizations were made within the overall quality categories. Thus, they were characteristics of papers at each level of quality but papers did not necessarily receive grades because they contained certain errors. Proofreading was neither encouraged nor discouraged. An overview of the analyses is given in Figure 23.

Analysis	Performed by	Attributes of the Essays Examined	Results of the Analysis
Overall quality	English teachers who were trained, experienced scorers.	Overall quality in each essay--no attempt was made to focus upon any particular writing attributes.	The results of the analysis were used to identify papers representing higher, middle, and lower quality. The essays identified were used in the characterization and computer analysis described below.
Error counts	Graduate students in English who had experience as English teachers.	Errors of nine different kinds were counted in each essay. The kinds of errors were: punctuation, spelling, capitalization, run-on sentences, improperly used sentence fragments, word choice, awkward construction, paragraphing, and agreement.	The number of errors provided some information about how children and young adults write. The error counts also pointed out some aspects of writing where children are in the greatest need of improvement, others where no remedial work is needed at all.
Characterization	Judges: English teachers, English scholars, and a linguist.	Higher, middle, and lower quality essays (as identified by the overall quality analysis) were examined to describe the varieties and natures of both the error and correct constructions, etc., in the essays.	The results of the analysis were statements about the essays at the three levels of quality. The statements provided a background for understanding the kinds of errors that occurred.
Computer	A computer using a program written for essay analysis.	The computer examined higher, middle, and lower quality essays (as indicated by the overall quality analysis) for the presence of particular attributes.	The information resulting from the computer analysis provided additional background for interpreting the error counts and the characterizations.

(Slotnick, 1972, p. 12)

Figure 23
An Overview of the Analyses Performed on the Essays.

Reporting

1. Results for each of the four analyses for each age group were reported separately.

2. A summary for each chapter recapped the major conclusions about the essays.

3. A table of summary statistics was presented separately for each age group. A sample of this table is Figure 24.

Factors which challenge the validity of the comparisons were discussed. An ordinal scale was constructed which showed how the higher, middle, and lower quality groups at each level related to each other. The lowest point on the scale was called the no mastery level, the midpoint was called the mastery of basics level, and the topmost point on the scale represented skillful handling of written language. There was no notion of goodness or badness built into the scale. Descriptions and examples of essays at each level were presented.

Reporting Categories	Analysis	Attribute	Essay Quality	Mean Errors Per 100 Words	Mean Errors Per Essay	% of Papers Having This Number of Occurrences: 0	1	2	3
Paragraphing	Error Counts	Number of Paragraphing Errors	High	+ *	+ *	98	2	0	0
			Middle	+ *	+ *	97	2	0	0
			Low	.1 *	.1 *	95	3	1	1
	Computer Analysis	Number of Paragraphs	High	.9	2.2	0	58	10	10
			Middle	.8	2.0	0	64	11	10
			Low	.9	1.6	0	78	12	5
		Mean Paragraph Length	High	**	10.1	Q_1 = 4.5		Q_3 = 16.2	
			Middle	**	9.4		4.6		14.0
			Low	**	6.4		4.2		8.8
Punctuation	Error Counts	Punctuation Errors	High	2.2	4.1	9	11	15	14
			Middle	2.9	4.8	18	9	12	16
			Low	3.5	4.4	7	17	9	13
	Computer Analysis	Number of Commas	High	2.0	4.8	11	12	11	11
			Middle	1.3	2.9	24	15	15	14
			Low	.9	1.3	49	20	9	8
		Number of Colons	High	+ *	+ *	99	1	0	0
			Middle	+ *	.1 *	96	2	0	0
			Low	+ *	+ *	99	0	0	1

* Single stars indicate arithmetic means, otherwise they are trimeans (see Tukey, J., Exploratory Data Analysis, Limited Preliminary Edition, Addison-Wesley, 1970, p. 2-25).

** Double stars indicate values not reported because of inapplicability.

\+ A plus sign indicates a value greater than .00 and less than .05.

When much variability appears in the data, quartile values are reported: Q_1 is the 25th centile, Q_3 is the 75th centile.

(Slotnick, 1972, p. 66)

Figure 24
A Portion of the Statistics for the Age 13 Famous Person Exercise Papers.

WRITING MECHANICS 1969-1974: A CAPSULE OF CHANGES IN WRITING MECHANICS

National Assessment of Educational Progress	Colorado 1975

Purpose

The major purpose of this project was to evaluate and analyze the characteristics of the essays written by two equivalent samples of students at three age levels. The age levels used were nine, thirteen, and seventeen years.

Methodology and instrumentation

1. For each age level the sample of 1969 papers was randomly mixed with the sample of 1974 papers.

2. The papers were then scored holistically as described in the previous summary. This provided a reliable ranking of the essays.

3. Another group of teachers scored each essay according to a specific scoring guide. These readers coded each paragraph and sentence for its type and coded each mechanical error found as described in the previous summary.

4. The essays and codes were keypunched and the results were computer-tabulated as described in the previous summary.

The two scoring methods were used in order to characterize specific writing accomplishments better than either approach could do alone.

Analysis

Each age level was reported separately. Main areas that were considered in the report were profile of the average essay written in 1974, changes in the average paper from 1969, range of the essays, and male-female differences.

Major outcomes and applications

1. The 1974 average essay for thirteen and seventeen-year-olds was of lower quality than the 1969 essay.

2. The 1969 and 1974 essays of nine-year-olds were similar in quality.

3. This procedure allows for valid comparisons to be made between essays written in different years and therefore, the identification of trends in writing.

4. Approximately one half of the exercises used in the 1973-74 assessment have been released. Each exercise was accompanied by the appropriate documentation and individual primary trait scoring guide. A "Writing Mechanics and Grammar Scoring Guide" was included. Tables presented for use with the released 1973-74 exercises included number of writing exercises by age level, number of writing exercises by age level showing the type of scoring, number of writing exercises by age level showing administration times and type of scoring for the three major objective areas, and listing of released writing exercises.

THE PRIMARY TRAIT SYSTEM FOR SCORING WRITING TASKS

Ina V. S. Mullis Colorado
1976

The rationale underlying primary trait scoring was that writing is done in terms of an audience and can be judged in view of its effect upon that audience. The assigned primary trait score essentially indicated whether or not a sample of writing contained the traits it needed in order to accomplish its purpose. The scoring procedure was based on the premise that a carefully constructed testing situation should provide an opportunity for respondents to demonstrate their ability to choose and carry out effectively, appropriate rhetorical strategies. An assessment of this writing had to be based on its likelihood of achieving the desired effect. The features that contributed to this success must be identified and defined in terms of their importance. These definitions then became the scoring criteria.

The three things specified for each exercise were the identity of the writer, the audience, and the subject matter. Each exercise was constructed to have some motivating feature. The scoring of each exercise was done on a four-point scale. Usually one indicated absence of the primary trait, two indicated presence of the primary trait, three indicated competence of the primary trait, and four indicated excellence of the primary trait.

Details were presented in Mullis's initial publication for scoring primary traits found in three types of writing - explanatory, persuasive, and expressive. Information was also given on scoring papers with respect to several secondary traits. Each paper was scored by two readers and reconciled by a third. Procedures for training readers were suggested. Examples of detailed primary trait scoring guides for particular exercises were given.

EXAMPLE OF A LOCAL EDMONTON PUBLIC SCHOOL DISTRICT APPLICATION OF THE ASSESSMENT OF WRITING UTILIZING THE PRIMARY TRAIT SYSTEM

Alberta
1976

Background

This example of a scoring guide was developed as part of a Language Arts assessment development project which was conducted by the Edmonton Public School District from September 1974 to June 1976. This project was coordinated by the Department of Research and Evaluation, Pupil Assessment Branch in cooperation with the Curriculum Department of the Edmonton Public School Board. The development of this scoring guide illustrated an application of the Primary Trait Assessment procedure to a junior high Language Arts objective currently used in the Edmonton Public School District.

Identification of primary and secondary traits

The primary trait for a plotted narrative as defined by Mullis (1976) was the presence of a properly developed and resolved conflict. Some secondary traits which may be considered in addition to the primary trait were point of view, use of dialogue, and presence of fantasy.

The four levels of competency of a primary trait could be defined for all plotted narratives. However, each writing task developed would need a scoring guide specifically tailored to the given task.

In general, the following definitions could be used to score a plotted narrative writing. The Edmonton Public School District developed a sample of such a scoring guide to score a piece of plotted narrative writing.

Sample scoring guide

The scoring guide was developed according to the previously defined criteria. Scores of four, three, two, and one were defined. Suggestions for scoring secondary traits were given. The next section of this summary illustrates this scoring guide for assessing the presence of a properly developed and resolved conflict, which is one primary trait of a plotted narrative.

For a score of four to be awarded which represents excellent or outstanding work, a properly developed and resolved conflict must be presented. In these papers, four basic elements must be evident.

First, the problem or setting must be presented. This element introduces the character in a certain place at a certain time. Development of the setting basically involves descriptive writing where necessary impressions or moods are established. Included here are details of the people and places involved in the story. The character can be animal, vegetable, or mineral; the place could be earth or not earth, air, land, or sea, city or country, indoors or outdoors, etc.; and time as distinct from tense could be past, present, or future and could include time of day. The second necessary element is the rising action. In this section, the hero's goal or conflict must be established. Appropriate and varied complications to the hero's achievement of his goal are introduced. Various types of complications are: hero versus self; hero against man; hero against group, family, town, society; hero against environment; hero against beast. Development of the climax is the third element. The hero either wins or loses, reaches or does not reach his goal. The fourth element is the falling action. This element may be very effectively omitted by some creative writers. Generally however, an appropriate ending, for example, logical or illogical, fair or just should complete the narrative. In addition to the presence of these four elements, appropriate sophistication of the composition suitable to the grade level must prevail throughout the composition. Included here would be such things as the nature of the complications introduced, that is, their number and complexity, the richness and vividness of descriptions used throughout the composition; creation of the right mood for the story; and level of vocabulary used.

For a score of three to be awarded which represents competence of work, compositions will be technically correct but have a lesser level of originality, sophistication, and creativeness. Alternatively, the composition could be deficient in some structural aspect, for example, in the length or number of complications presented, but if the high level of originality and creativity is evident, the composition would still rate a three.

For a score of two to be awarded which represents presence of the primary trait, compositions are correct in the basic format of the narrative but deficient in grammar, style, or syntactical structure. The basic sequencing of the elements of the story may be correctly ordered but the level of sophistication is inappropriate.

A score of one is awarded where there is an absence of the primary trait. Structural defects as well as technical or grammatical deficiencies or omissions are evident in these compositions. Elements in the narrative are not properly developed. This could be manifested by placing the conflict in the setting, making complications into events rather than stumbling blocks for the hero, developing a climax which is unrelated to or does not solve the problem, or failing to make the hero's goal active. An example of the latter error would be centering the story around "the hero is in a forest fire" rather than "the hero's goal is to get out of the forest fire."

Additional definitions for each secondary trait could also be developed for the plotted narrative. Scoring here is usually based on presence or absence of the trait or on distinct differences within a trait. An example of the latter situation is as follows: (1) a code of one will be assigned if the point of view is consistently one of the participants in the story, for example, "If I were..."; (2) a code of two will be assigned if the point of view is consistently one of the observer; and (3) a code of three will be assigned if the point of view cannot be determined or is not consistent.

Similar scoring guides for the description, exposition, and argument types of writing could be developed. Refinement and tailoring the definitions to each particular writing task would then be relatively straightforward.

TOWARD A THEORY OF COMPUTER ESSAY GRADING

Henry B. Slotnick Colorado
1972

Purpose

The major purpose of this study was to determine whether computer essay measurement could be grouped into factors.

Methodology and instrumentation

1. Three hundred twenty-six papers were used from an original set of four hundred seventy-six high school student papers written on one topic.

2. The essays were keypunched and processed using the computer program Son of MOMSR. Thirty-four measurements were made on each paper.

Analysis

1. Means, standard deviations, and correlations were calculated and a principal components factor analysis was conducted.

2. Six factors were identified: fluency, spelling, diction, sentence structure, punctuation, and paragraphing. Factor scores were computed for the original essays and two sets of papers were identified for each factor: high scores and low scores. These papers had average scores for all other factors.

3. The computer program had identified sets of papers which apparently differed significantly only in one of the factors. These papers were then analyzed in depth by readers to determine what attributes of interest were reflected by the factors.

Major outcomes and interpretations

1. The attributes found were: quality of thought, spelling, range of vocabulary and word choice, structure of sentences, emphasis through subordination, and paragraph organization.

2. A general model of computer essay grading has been identified which allows for computer tabulations of very detailed measurements of essays, provided an IBM 360 is available. These procedures have since been adapted by National Assessment of Educational Progress (NAEP) in their writing assessment.

APPENDIX C:

SUMMARIES OF RESEARCH RELATED TO ASSESSMENT OF FUNCTIONAL LITERACY

FUNCTIONAL LITERACY, TECHNICAL SUMMARY

Charles Gadway
H. A. Wilson

Colorado
1975

Purpose

The purpose of the Mini-Assessment of Functional Literacy (MAFL) was to assess the extent of functional illiteracy of high school graduates in the United States.

Methodology and instrumentation

1. MAFL exercises were selected to present the formats of reading materials frequently encountered in everyday life.

2. These eighty-six exercises were compiled into two booklets in five formats, each of which could be administered in one hour. The formats used were passages; drawings, pictures, coupons; charts, maps, graphs; forms; and reference materials.

3. These exercises were administered only to seventeen-year-olds enrolled in school in 1974.

Analysis

1. Scoring, data processing, and analysis were done in conjunction with scoring and analysis activities for National Assessment of Educational Progress's regular reading and writing assessment.

2. Since sixty-four of eighty-six exercise parts had been included in the MAFL 1971 study, changes in performance were analyzed.

Major outcomes and interpretations

1. All groups gained in functional reading skills as measured by MAFL.

2. In general, those groups which had most to gain, gained most.

A CRITERION-REFERENCED ASSESSMENT OF READING LITERACY USING THE CLOZE PROCEDURE

Lee H. Hansen
Karl D. Hesse

Wisconsin
1974

Purpose

The major purpose of the study was to determine whether or not students in Madison Public Schools could achieve a correct replacement score greater than 35% on cloze tests developed from samples of reading materials which they were expected by the community and school to be able to read. Bormuth (1974) had established 35% as a criterion for minimal reading literacy.

Methodology and instrumentation

The following procedure was used.

A taxonomy of reading materials in ten domains was developed. Within each domain, a set of documented decision rules was used to obtain representative messages from each category. A random seventy-word passage was extracted from the message and set up as a cloze passage. Passages were administered to sixty students each at grades four, seven, ten, and twelve in the Madison Public School System. Between seven and eleven passages were included per booklet. Between one hundred forty-four and three hundred passages were tested at each grade level. The sample at each grade level was between 1080 and 2160 students. A multimatrix sampling design was used. A common ten-item cloze passage was administered to all students.

Analysis

1. Scoring was consistent with the Bormuth model.

2. A one-way analysis of variance was conducted among the subsamples at each grade level.

3. Percentages of students scoring above and below the minimal literacy criterion of 35% were calculated for each domain at a grade level.

Major outcomes and interpretations

1. It appears that substantial numbers (25-30%) of students leave high school unable to gain sufficient information from most basic materials society is expecting them to read.

2. This also appears to be true with respect to textbooks and

reference materials purchased for their use in school.

3. These results were found for a population who generally perform above national norms.

4. These results are consistent with the concept that the readability of commercially available materials is inappropriate.

SURVIVAL LITERACY STUDY

Louis Harris — New York 1970

Purpose

This study was designed to determine the percentage of Americans lacking the functional or practical reading skills necessary to survive in the United States.

Methodology and instrumentation

Interviews for this study were completed in over 1500 homes throughout the United States. A stratified random sampling technique was used to determine which state, county, city, town, and ultimately household would be approached. Each respondent was asked to complete five application forms used for obtaining a social security number, a personal bank loan, public assistance, medicare, and a driver's license.

Analysis

Scoring of application forms resulted in a raw score out of a total score which was then converted to a percentage correct. Data was keypunched and tabulated by a computer. Results were tabulated for percentage of total sample and several subsamples correctly responding to each item. Analysis further categorized results into four literacy groups.

1. Low Survival Threshold - Respondents averaged less than seventy percent correct. This group was considered functionally illiterate.

2. Questionable Survival Threshold - Respondents averaged less than eighty percent correct. This group filled out the forms with considerable difficulty and included members of the above group.

3. Marginal Survival Threshold - Respondents averaged less than ninety percent correct. Generally, this group lacked total survival reading ability and included members of the two groups above.

4. Likely Survival Threshold - Respondents averaged more than ninety percent correct. This group was considered to be fully literate for the purpose of this study.

Major outcomes and interpretations

1. Three percent of the sample fell into the low survival category, five percent into the questionable survival category, and thirteen percent into the marginal survival category.

2. In the United States practical literacy increased in direct proportion to income.

THE 1971 NATIONAL READING DIFFICULTY INDEX: A STUDY OF FUNCTIONAL READING ABILITY IN THE UNITED STATES FOR THE NATIONAL READING CENTER

Louis Harris

New York
1971

Purpose

The purpose of this study was to measure the ability of Americans to respond to practical, real-life situations by asking them to read and respond to different common printed material.

Methodology and instrumentation

A stratified random sampling technique similar to that of the Harris (1970) study was used to determine a national sample. Interviews were conducted with a total of 1747 adults age sixteen years and older. Each adult responded to several types of printed material. These included direct dialing instructions, classified employment and housing advertisements, and a variety of application forms. Materials were then subdivided into ten sections. Analysis scores were obtained for each section for the total sample and numerous subsamples. Subsamples included sex, age, race, educational background, geographical region, size of city, occupation, and place of birth. Raw scores out of the total possible score were converted to a percentage. All data was then keypunched and computer processed.

The authors then attempted to determine a National Reading Difficulty Index. Several possible procedures were considered before the index was calculated. It was decided to base the index on degree of difficulty of any item. Weights were assigned to each item based on the overall percentage who missed that item. This procedure for computing an index was selected in order to emphasize the area causing the greatest reading problem and to highlight differences in reading ability that exist among various subgroups in the United States. The Reading Difficulty Index was a scaled indication of reading difficulty (a score less than ninety percent). The higher the index, the higher the reading difficulty. An index of 100 indicated that the median score was below ninety percent correct.

Major outcomes and interpretations

1. Four percent of the population scored under eighty percent correct and were considered to be suffering from serious deficiencies in functional reading ability.

2. The reading index for the group earning under $5,000 was forty-eight while the indices of other groups ranged

from fourteen to twenty-six. The average United States index was thirty.

3. Reading indices decreased sharply with increase in educational background.

4. Procedures used in this study seemed to be in line with the views of the majority of researchers involved in the assessment of functional literacy. However, any application of this study to an adolescent population would require an examination of alternate materials more suitable to a younger population.

THE DEVELOPMENT AND VALIDATION OF R/EAL: AN INSTRUMENT TO ASSESS FUNCTIONAL LITERACY

Mary Lichtman

District of Columbia
1973

Purpose

The major purposes of this study were:

1. to provide information about the design and development of Reading/Everyday Activities in Life (R/EAL) including determination of reading criteria, establishment of task analysis, production and selection of individual test items, development of test format and procedure and completion of the final version of R/EAL;

2. to provide information about the validation of R/EAL.

Methodology and instrumentation

During initial preparation and construction of the instrument, material prepared had to fit a three-point criteria. The materials must be representative of activities directly related to practical life reading experiences, be suitable for adolescents and adults, and be presented so as to resemble the appearance of the material as it is usually found in real life situations. The test was constructed following basic criterion-referenced test development procedures. An individually controlled audio cassette was used for all directions and information. An open-ended response mode rather than multiple-choice format was selected for the test.

Field testing was conducted during 1971-72 on 300 inner-city students from the Washington, D. C. area. The test was analyzed and was revised to consist of forty-five items divided into the following nine categories: (1) signs and labels, (2) schedules and tables, (3) maps, (4) categorized listings and indices, (5) high interest, factual, narrative (6) illustrated advertisements, (7) technical documents, (8) sets of directions, and (9) fill-in the blank form. Procedures were implemented to establish validation of R/EAL. Four hundred thirty-four subjects aged sixteen to twenty-one enrolled in a residential manpower training program were given the R/EAL. Each student had his own individually controlled and operated cassette recorder and earphone, and was permitted to work at his own pace.

Analysis

Item analysis, factor analysis, reliability values, and validity data were obtained for the test.

Major outcomes and interpretations

1. Results from this study support the use of R/EAL as a viable assessment instrument for functional literacy.

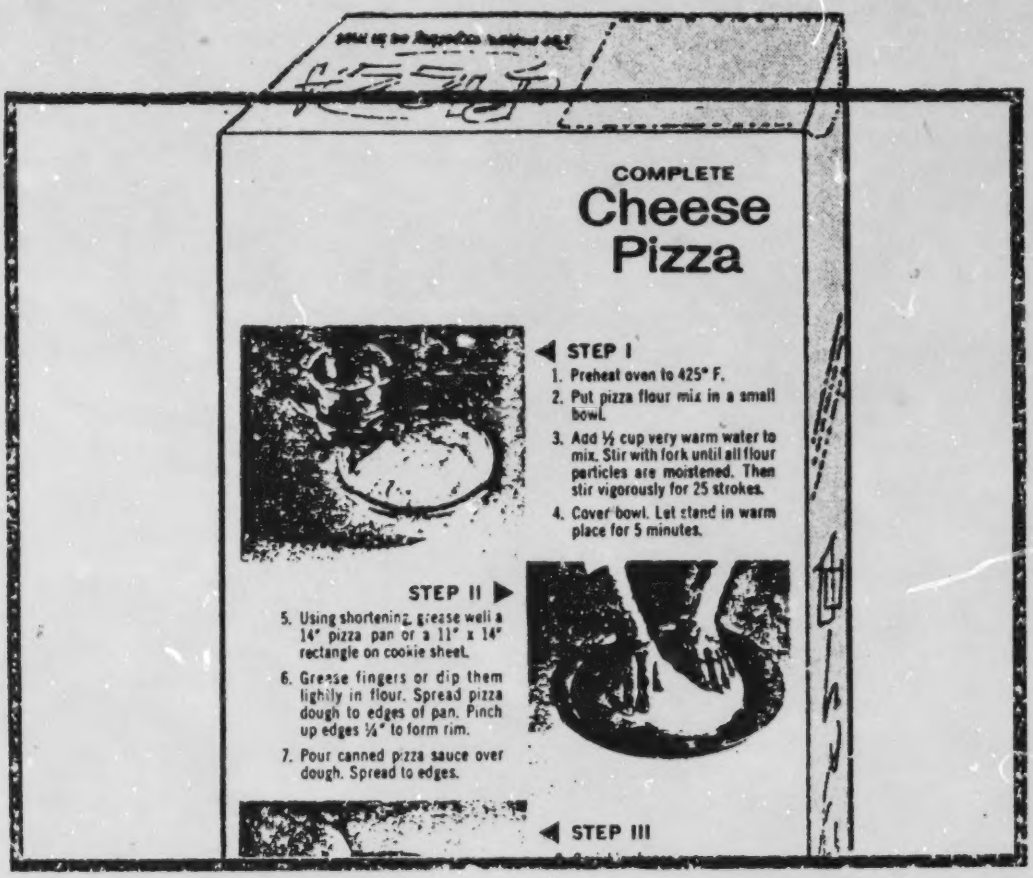
2. A larger variety of functional literacy tasks were incorporated into this test than for similar tests in the past.

Sample

SUBTEST 3 - DIRECTIONS FOR FOOD PREPARATION

The reader is presented a set of directions for preparing food. He is required to interpret the parts of the directions in a manner that will result in a satisfactory finished product.

Example: What must you do before you mix the ingredients?


COMPLETE
Cheese
Pizza
STEP I
1. Preheat oven to 425° F.
2. Put pizza flour mix in a small bowl.
3. Add ½ cup very warm water to mix. Stir with fork until all flour particles are moistened. Then stir vigorously for 25 strokes.
4. Cover bowl. Let stand in warm place for 5 minutes.
STEP II
5. Using shortening, grease well a 14" pizza pan or a 11" x 14" rectangle on cookie sheet.
6. Grease fingers or dip them lightly in flour. Spread pizza dough to edges of pan. Pinch up edges ¼" to form rim.
7. Pour canned pizza sauce over dough. Spread to edges.
STEP III

(The remainder of the page was devoted to space for the student to answer the question.)

(Lichtman, 1972, p. 21)

Figure 25
Sample Functional Literacy Question from R/EAL.

CRITERION-REFERENCED TESTING FOR FUNCTIONAL LITERACY

Thomas F. McDonald
Gary B. Moorman

Arizona
1974

Purpose

Tha major purpose of this study was to develop a district-wide criterion-referenced reading test which would demonstrate the proficiency of high school graduates to read at least at the ninth grade level.

Methodology and instrumentation

The following sequence of events occurred.

1. A district committee of reading specialists was formed in September 1971.
2. Objectives were developed in basic reading skills which an adult in our society needs in order to function without handicaps.
3. After a year of review by various other publics, the objectives were revised to twelve performance objectives for the Minimal Reading Proficiency Assessment (MRPA).
4. A second committee developed over 150 test items.
5. These items were field tested on a random sample of 1000 students within a Phoenix school system. Each student wrote one of three forms of fifty items.
6. A third committee selected four final items for each of the twelve objectives. Each item was in multiple-choice format with five alternatives. Items were selected on the following three criteria: an item had to relate to the students' interests, have a difficulty level of approximately .75, and effectively discriminate between an adequate and an inadequate reader.
7. In November 1972, the MRPA was administered to 12,170 freshmen and sophomores in the district. Passing requirement was 75%.

Analysis

1. The number of students and skill areas passed and not passed were tabulated for each of freshmen, sophomores, and total.

2. The overall average difficulty as well as average difficulty for each of the twelve areas were calculated.

3. The correlation between MRPA and the Davis Reading Test was calculated and found to be 0.89.

Major outcomes and interpretations

1. Two additional forms of MRPA were subsequently developed.

2. Study skill packets, computer skill packets, and banks of reading materials and exercises for each of the twelve areas have been developed.

3. All students in the Phoenix Union High School System now must pass all twelve skill areas with 75% proficiency in order to graduate from high school.

4. There still remains the problem of developing a uniform criteria for classifying items into skill areas.

5. There is a limited number of questions in each skill area.

ADULT FUNCTIONAL READING STUDY

Richard T. Murphy
Educational Testing Service
New Jersey

Final Report December 1973
Supplement December 1975

Introduction

This project was undertaken to provide information on the reading activities and skills of American adults as they go about their ordinary activities in the course of a normal day. This project involved:

1. a national survey (N = 5096) of adult reading activities in order to identify the reading tasks actually performed by adults in a regular day;

2. the construction of a set of reading tasks intended to sample the universe of reading tasks ordinarily encountered by the typical adult in our culture;

3. a national survey (N = 7866) of actual adult performance on this set of reading tasks.

In this project, no attempt was made to define functional literacy. However, in identifying reading tasks performed by adults in carrying out their ordinary activities it is assumed that such tasks will have a favorable return to the individual and to society in general.

Identification of reading tasks

Interviews conducted in the Reading Activities Survey were based on a structured questionnaire designed to help the respondent recall the general daily activity he was involved in the day before the interview, the specific type of reading which occurred in each of the general activities, the time spent in each activity, and the importance of the activity.

Questions reflecting thirteen socio-economic variables were also included. Following the survey it was thus possible to obtain this type of answer to the question, "What do adults read?" The following types of materials were included in the answers given. The reading of mail was reported by 53% of the respondents. Bills or statements were considered most important and were read by 26% of the total population (or 49% of those reading mail). Bill or statement readers were of average socio-economic status (SES). Advertisements, although read by a similar percentage of the population were considered least important of the mail received.

The Reading Activities Survey indicated that the average person read about ninety minutes in a typical day. Most of this time was spent on reading newspapers, magazines, books, and job related matter. Thirteen such general activities were identified and these, in turn, were further broken into 132 specific reading activities. The proportion of the population engaged in each activity in a typical day was indicated along with their SES and an index of their stated importance of each task as shown in Figures 26 and 27.

Five percent of all adults could not read in the English language. These persons included the visually handicapped, the foreign language readers, and the illiterates. The inability to read, however, should not be confused with "functional illiteracy." The second and third phases of the project attempted to measure, "What are Americans able to read?"

Development of the reading tasks

The objective of this project was to develop reading tasks that would reflect as faithfully as possible the characteristics of reading tasks as they are generally encountered in real life. The tasks developed were not subject to the usual constraints of conventional test items. No separate answer sheets were used. The only constraints were that the tasks be life-like, that they be beneficial to individuals who could perform them (tasks were expected to discriminate adults who could function independently in society at a minimal level from those who could not), and that a successful response to a task be observable and scorable.

Reading tasks were classified according to (1) the general form of the material to be read and (2) the benefit area (describing the kind of return that accrues to an individual who can perform the task, for example, economic benefit, occupational, educational, or recreational benefit).

In the first round of development, several hundred real-life tasks (for example, a real medicine label rather than a regular sheet of paper with the label typed on it) were constructed. The tasks were then submitted to a panel of persons outside Educational Testing Service (ETS) for criticism, direction, and judgment of the benefit classification mentioned above. The final result of the ensuing revisions was a set of 269 reading tasks arranged in 184 items. Traffic signs, product instructions, periodical items, recipes, correspondence, legal forms, and advertisements were some of the tasks included. For field testing purposes, the remaining tasks were distributed and arranged according to an item sampling procedure into seven different sections. Each section contained from thirty-two to forty-five tasks of varying degrees of difficulty and representative of each benefit category and form. Each of the task books was

administered to approximately 300 respondents. Using the results of the field testing phase, the final set of 170 reading tasks was then arranged into ten booklets of seventeen items each for the national reading survey. The booklets were arranged to balance for benefit type, general form, and difficulty level as shown in Figure 28.

The National Reading Performance Survey

Seven thousand eight hundred sixty-six persons were interviewed in the survey. Specific sample locations, households and individuals were designated by a probability sampling model. All interviewers received formal training in strict probability sampling procedures.

The survey instruments for each respondent consisted of a brief demographic questionnaire administered orally and the reading task book. The interviewer read the directions for each task and the respondent attempted to mark his answer in the task book. The average time spent by an individual on a task book was less than half an hour.

The reading tasks were then carefully coded by ETS staff as correct, incorrect, or omit. Several weighting adjustments were then made to the data in order to correct for sampling bias and thereby make the data more representative of the general population.

Correct response percentages were obtained on each of the 170 tasks for the total population (N = 7866) and for approximately fifty groups. Each reading task was attempted by approximately 750 persons.

It was thus possible to ascertain actual performances on each of the 170 reading tasks. In this particular survey, twenty percent of the reading tasks were answered correctly by more than ninety percent of the population. Almost half of the items were answered correctly by more than eighty percent of the population. In addition to the examination of individual reading tasks, it would be possible with input from reading specialists and other experts in the field of adult literacy to examine groups of items in an attempt to develop combined measures of reading competency.

Supplemental study

As a supplement to the original study, the tasks below were carried out during the period January 1974 to December 1974. An analysis of errors committed by respondents in the National Reading Performance Survey was made. An inquiry into the reasons why certain errors were committed was made.

In order to analyze the kinds of errors made by respondents, a ten percent random sample of respondents was chosen and their responses examined in considerable detail. From this analysis, approximately 900 distinct errors were identified. Two very simple

causes of incorrect responses emerged. It was found that many persons did not understand the individual words in the reading tasks. In addition, some did not understand how to extract information from the forms, charts, or tables in which such words appear in such items as doctor's bills, train schedules, application forms, or election forms.

A study of the relationship of the reading tasks to decoding skills was made. In this study, thirty-three of the adult functional reading items from the National Reading Survey were administered to seventy-seven adults having a wide range of decoding abilities, together with a set of materials developed by ETS designed to assess decoding skills. The purpose of the study was to examine whether or not the inability to answer correctly the adult reading tasks was attributable to inadequate decoding skills. The first two parts of the test (Alphabet Test and Test of Word Attack Skills) were group administered; the Test of Oral Reading was individually administered. None of the tests were timed. Each item was scored on a pass-fail basis according to a detailed scoring guide.

The difficulty levels on the thirty-three tasks were calculated as well as the reliability which was a KR_{20} of 0.94. Difficulty level, percentage correct (or passing) and reliability were calculated for each item on the thirty-three item Survey of Reading Material as well as on the Test of Word Attack Skills portion of the decoding instrument. Mean levels of difficulty for the Test of Word Attack Skills and Test of Oral Reading were charted and intercorrelations for ten major measures were calculated.

As was expected, the data from this study clearly indicated that the ability to decode is a predictor of success on the sample of adult reading tasks. However, it was not possible to determine a critical score on various decoding measures which would predict this success. Further research is needed to identify a critical level of decoding ability for adult functional literacy.

A study of the relationship of the reading tasks to a set of cloze tests was completed. In this phase of the study, scores on functional reading tests and scores on cloze tests were related in an attempt to demonstrate concurrent validity between those two methods of measuring reading skill. In a close test, every n^{th} word is deleted from an appropriate passage and the subject is asked to "fill in the blanks." The cloze tests used were based on a subset of thirty-six passages that had previously been scaled for difficulty using an extensive set of tests. The reading tasks were a subset of the original 170 functional reading tasks previously developed by ETS.

Since each reading task was scored as either right or wrong and each cloze test could produce scores from zero to thirty (one point for each deletion correctly replaced) a cutting point of eleven was selected as that score at or below which the passage would

be scored as "wrong" and above which the passage would be scored as "right." The percentage of the population scoring "right" on the reading tasks could thus be directly compared to the corresponding group on the cloze passage.

Two tests for each of grades seven, ten, and twelve were then constructed. Each contained eighteen reading tasks and three cloze passages. The tests were then administered to 787 students in thirty-two classes in seven schools. Average administration time was forty-five minutes. The means, standard deviations, percentage passing and reliabilities for each of the reading tasks and cloze passages were calculated. Intercorrelations of the total reading and cloze scores for each of the two administration days were also calculated. Reading tasks were then ranked according to difficulty level and specific cloze passages were thus matched at approximate levels. Because the reading tasks were generally easier than the cloze passages, in some instances a cloze passage ranked below all the reading tasks. If students could read such a cloze passage they would probably read most of the reading tasks above it.

Tentative conclusions reached by the study indicated that within a grade level, scores of the students on the reading tasks and the cloze passages were moderately correlated. For example, in grade ten, nearly sixty percent out of a possible seventy-eight percent of the variance in the cloze tasks was predictable from reading task performance. It thus appears that students who can complete certain passages can probably pass certain functional reading tasks. In addition, however, at each grade level students who failed to meet the thirty-five percent criterion (eleven replacements) on cloze passages of moderate difficulty did manage to pass a large number of adult functional reading tasks.

Since the functional reading tests were considerably less difficult than the cloze tests, further research using cloze tests developed on more functional kinds of reading passages would prove useful.

Development and field testing of an experimental test of basic reading competency in the schools was done. This phase of the study attempted to answer the question, "Are the reading tasks developed in Adult Functional Reading Study suitable for assessing the reading skills of school students?"

One hundred twenty of the one hundred seventy reading tasks were initially selected for trial. After review and revision, eight test forms with fourteen or fifteen tasks on each form were developed and balanced for topic, skill, difficulty level, benefit, and form. Initial field testing on 2033 grade nine students resulted in an average item difficulty of 0.80. Further revision of items resulted in two sets of forty items each being field tested on another 2520

grade nine students. Mean difficulties were 0.76 and 0.78. Reliability of Form A was 0,85 and of Form B was 0.84, The inability of the students to finish the test indicated a shorter test or longer administration time was needed.

In general, the reactions to these materials were favorable. Teachers thought the materials were reasonable for assessing functional reading. Students thought the materials were interesting and were motivated to respond to them. It thus appears that the reading tasks in the experimental test have a certain face validity. But whether or not students who can respond correctly at some cutoff score are in some way sufficiently prepared for surviving in a real-life reading world is still an unresolved question.

Samples

General Activity[a]	Percent Readers	Average Reading Time (minutes)	SES	Correlation of Reading Time with SES
Newspapers	73	35	Average	.19
Magazines	39	33	High	.19
Books	33	47	High	.17
Mail	53	5	Average	.15
Meals	42	3	Average	.06
At Work	33	61	Average	.19
Working Around House	46	7	Average	.10
School	5	68	High	.07
Traveling	70	3	Average	.07
Shopping	33	7	Average	.07
Club, Church Activities	10	16	Average	.04
Theater, Game, or Event	4	7	High	.05
Recreation, Free Time	54	7	Average	.07

[a]Categories are mutually exclusive.

(Murphy, 1973, p. 22)

Figure 26
Reading Occurring During General Daily Activities.

General Activity	Percent Readers	5 Minutes or More	SES	Importance
Main News	66	52	Average	43
Local News	55	41	Average	33
Editorials	38	25	High	18
Women's Pages	37	21	Average	9
Sports	36	24	High	17
Comics	36	18	Average	6
Regular Ads	35	15	Average	10
Classified Ads	29	17	Average	12
TV-Radio Listings	27	8	Average	7
Financial News	22	12	High	13
Movie, Book Reviews	16	7	High	3
Magazine Section	12	9	Average	4
Obituaries	7	3	Average	4

(Murphy, 1973, p. 25)

Figure 27
A More Detailed Sample Breakdown of the Newspapers Category Found in Figure 26.

Item Number	Description	Benefit	Form
1	List of book titles	Education/culture	Book
2	Signs on doors	Maintenance	Miscellaneous
3	Birth announcement	Personal relationships	Personal communcation
4	TV Schedule	Recreation	Listing
5	Anti-rabies form	Health	Form
6	Cereal coupon	Economic	Advertising
7	Military enlistment ad	Occupational	Advertising
8	Legal form	Occupational	Legal documents
9	Packet of seeds	Recreation	Set of instructions
10	Group insurance plan	Health	Miscellaneous
11	Table of contents	Maintenance	Listing
12	Airplane information	Health	Set of instructions
13	Train schedule	Maintenance	Listing
14	Postal rates	Citizenship	Set of instructions
15	Vacuum cleaners	Economic	Periodical
16	Passage on antelope squirrel	Education/culture	Periodical
17	Election ballot	Citizenship	Set of instructions

(Murphy, 1973, p. 79)

Figure 28
Arrangement of a Test Booklet by Benefit Type, and General Form.

FUNCTIONAL LITERACY FOR ADULTS: A STATUS REPORT OF THE ADULT PERFORMANCE LEVEL STUDY

Norvell Northcutt

Texas
1974

Purpose

The purposes of the Adult Performance Level (APL) study were to describe adult functional literacy in pragmatic, behavioral terms and to develop devices for the assessment of literacy which would be useful at a variety of operational levels.

Methodology

The APL study developed a system of adult literacy objectives and accompanying test items which were keyed to four literacy skill areas (communication, computation, problem solving, and interpersonal relations) and five general knowledge areas (occupational knowledge, consumer economics, community resources, government and law, and health).

Analysis

A complex sampling design by county was used to select the national sample who which was tested in the areas of occupational knowledge and consumer economics. All information was collected by interview and all data was keypunched and computer processed. Proportions of the population who correctly answered each item were tabulated.

Major outcomes and interpretations

1. Results suggested that far more adults were illiterate in the sense of being unable to apply skills to problem areas encountered in everyday life than would be expected.

2. An example of the results indicated that seventeen percent of the sample was not able to determine which one of four advertisements was placed by a private person rather than a corporation or private institution.

3. A second example indicated that twenty-two percent were unable to address an envelope well enough to ensure that it would reach its destination.

APPENDIX D:

SUMMARIES OF RESEARCH RELATED TO THE ASSESSMENT OF ORAL LANGUAGE

ORAL LANGUAGE: SPOKEN ENGLISH
CHAPTER 10 OF A LANGUAGE FOR LIFE

Sir Alan Bullock Great Britain
1975

Purpose

The purpose of this report was to discuss the status of language arts in Great Britain and make recommendations for future practices in the subject area. This summary will present some of the major points presented in the report's section on Spoken English.

Report content

During the past ten years there has been a substantial increase in the teaching and examining of spoken English. Different types of oral examinations are given in individual districts. The most commonly represented components are in prepared talk, conversation and reading aloud. Certain school boards have included in the course work activities such as improvised and scripted drama, debates, tape-recorded interviews and aural comprehension. The only General Council Examinations (G.C.E.) "O" level examination in Spoken English at present available is that of the London Board, which gave as its aim, "to test articulation and fluency." Its requirements are reading aloud and group and individual conversations with the examiner.

Teachers have supported the oral language emphasis in various degrees. Some have been concerned that testing reflected an artificial emphasis and did not reflect the main objectives of the Spoken English program. The report emphasized the positive aspects of teaching and testing Spoken English while recognizing the many concerns of individuals working in the area. The oral communication section closed with recommendations for better teaching and testing in Spoken English.

Outcomes and interpretations

1. The report emphasized support of oral language and its subsequent measurement. The author indicated that in today's society, talk is taking on an ever growing significance.

2. The report stressed the need for meaningful natural testing situations to validly assess oral communication.

163

A TEST USING ELICITED IMITATIONS IN ASSESSING GRAMMATICAL STRUCTURE IN CHILDREN

E. Carrow

Texas
1974

Purpose

The purpose of this project was to develop an instrument and procedures to measure the developmental status of a child's base sentence construction and grammar. The procedure was designed to provide a reliable but efficient method of obtaining performance data on a child's grammatical system. It included tasks of a wide range of complexity covering many verbal speech forms; permitted the examination both of what the child said as well as permitted comparison with adult speech; and was administered on an individual basis.

Methodology and instrumentation

The test, Elicited Language Inventory, consisted of fifty one sentences and one phrase of various lengths and degrees of complexity. It was administered to each of several hundred children over a two-year time span by recording on audiotape equipment, the child's imitations of the stimulus sentences of the examiner. These responses were then coded on a separate sheet into set grammatical forms.

Analysis

All coded responses were scored for two error types - category errors (plural, verb, negative, and so on) and by type errors (omission, substitution, transposition, and so on). Errors were defined as different from adult language even though the "errors" may not have been inappropriate for a specific age. Calculations included total error score and subscores for each category and type. These scores could be compared against norms.

Outcomes and interpretations

1. The Elicited Language Inventory seemed to give information similar to that obtained using a sample of spontaneous language.

2. The instrument provided reliable and efficient information about error patterns the child might have.

3. In its current form it is most useful in identifying, understanding and noting specifications of language disorders in children. However, adaptations of the instrument would be very useful in assessing normal children.

HYPERCORRECTION IN TEST RESPONSES

Courtney B. Cazden Massachusetts 1975
California 1975

Purpose

This paper discussed some of the difficulties in testing oral language with primary children when a standardized instrument was used rather than a more natural setting. The author promoted the use of condensed forms of familiar interaction experiences with a small group of children as the best way of obtaining a measure.

Methodology and instrumentation

Both subtests of CIRCUS and condensed forms of familiar interaction were tried with a group of seven California first and second graders who were native speakers of English. The children were tested on their oral language skills in three areas - the descriptive use of language, the functional use of language, and the narrative use of language. The results of both testing forms were tabulated and analyzed.

Analysis

On the CIRCUS test the children made thirty-five overgeneralizations out of fifty-six responses dealing with irregulars. Using the preferred second form of testing for those aspects of irregulars that could be tested that way, errors dropped sharply. The comparison of these scores to actual classroom spontaneous oral speech seemed to agree much better.

Outcomes and interpretations

1. The standardized oral production test produced more student errors than the less structured student testing situation. A possible explanation is that the more formal testing situation produces an unnatural set in speech patterns, particularly among young children when they are faced with many examples of a particular speech form at one time.

2. The author felt a more extensive comparison of the two test types would indicate other possible measurement distortions in the formal testing situation.

3. In general, this article pointed out the need for natural testing situations with young children.

HOW WELL CAN OUR STUDENTS SPEAK?
(LANGUAGE LABORATORY TESTING)

A. E. Deyes

Spain
1971

Purpose

The purpose of the study was to develop and test an instrument that could be used to assess students' oral language ability in a language laboratory.

Methodology and instrumentation

Three different tests were administered. The third test was developed following the conclusion that tests one and two were unsatisfactory. Test three was administered to university students and consisted of five parts. These parts were: recognition of sound segments; production of sound segments; intonation; structures; and oral compositon testing the three elements.

There were forty questions testing twenty patterns. Each question was repeated twice before the student was required to answer. In the fifth section of the test, students had to fill in gaps in a prose passage, using their own ideas. They were given two minutes to prepare the passage but were not allowed to write anything in advance or during presentation.

Analysis

All sections of the test were marked objectively. Analysis showed a reliable KR_{20}.

Major outcomes and interpretations

1. The study concluded that a satisfactory language laboratory test of oral production had been devised for any level of student.

2. The test was relatively simple to prepare, administer, and only took an average of twelve minutes marking time per student.

3. The testing situation has applications for English as a first or second language.

AN OBJECTIVE TEST OF PRONUNCIATION FOR LARGE CLASSES

Michael Dobbyn 1976

Purpose

The author developed an objective test of pronunciation that was quick to administer and yielded reliable results in a regular classroom situation.

Methodology and instrumentation

The principle employed in the test was that of preventing the examiner from knowing in advance what the student was trying to say. Cards testing the required pronunciation were prepared in random order so the teacher had to rely on the student's pronunciation to determine what was being said. The student read five of these prepared sentences and the teacher scored his pronunciation according to which sentence he thought it was as the student read. The average testing and scoring time per student was one minute.

Analysis

Depending on the content of the sentences, the tester (teacher) obtained diagnostic information about a particular set of pronunciation habits of the students. These results were tabulated for maximum usefulness.

Outcomes and interpretations

1. Using this method teachers would be able to collect required information about oral communication skills and thus structure their program about their findings.

2. The ease of construction, administration, and scoring makes this a very handy tool for the classroom teacher.

TOWARDS OBJECTIVITY IN GROUP ORAL TESTING

David Folland
David Robertson

Finland
1976

Purpose

This article was intended to suggest a possible way of solving the problems of objectively assessing oral conversation. Assessment problems were considered from both the viewpoint of English as a second language and a first language.

Procedures and methodology

The authors described the history of their test development program, indicating the advantages of their oral testing procedures over the traditional interview or similar unstructured measurement procedures for university aged students. Their group test consisted of groups of students, limited to a maximum of seven, who had attended the same lecture course and discussion group, being tested in free discussion by two English teachers, for a minimum of five minutes per student. The setting for the test was as unlike a classroom as possible and a pre-recorded tape was used at the start as a focus for the discussion. The two English teachers did not participate in the discussion but instead scored the students as they contributed to the discussion. The student contributions to the discussion were unrehearsed. Marks were awarded for pronunciation, lexis, grammatical structure, use and number of utterances (one utterance being approximately ten words or more) up to ten. Errors in the four categories were ranked on a four-point scale according to degree of seriousness.

Analysis

All student oral scores were combined in this application with their reading and comprehension scores for a total student mark. The marking structure outlined above has been analyzed for reliability and found to be highly reliable.

Outcomes and interpretations

1. The authors believed that the group test and this marking system could be usefully and easily adapted for use in other situations and for other levels of language ability.

2. The marking system is very flexible and can easily be adjusted.

3. An important use of this technique in language arts courses would be the objective assessment of conversational skills.

TOWARDS A TEST OF LANGUAGE PRODUCTION

Norman Graham Great Britain
1971

Purpose

The major purpose of the study was to develop a test that would measure oral language production in primary school children.

Methodology and instrumentation

Instead of measuring spontaneous utterances in normal surroundings, it was decided to place a subject in a situation which would maximize the chances of his producing a particular structure in an utterance if it was within his competence to do so. The principle used here was maximum cueing short of mere repetition.

The test consisted of twelve items each manifesting a particular grammatical transformation. Each item consisted of six pictures which were used to elicit a particular sentence type.

After practise, each child was shown a picture and asked to verbalize the passage the picture portrayed. If the correct structure was included in the response, the next picture of the series was shown. If the structure was not in the response, the examiner gave the correct message, asked the child to repeat it and then went on to the next picture.

Analysis

Each item was scored correct if any one or more pictures elicited the desired response.

Major outcoomes and interpretations

1. Preliminary results with educationally subnormal and normal nursery school children suggested that the children tested know much more about their native language than superficial observation would indicate.

2. Children appeared to be learning from the examples given if their responses were incorrect. They were learning precisely what was required of them for that particular situation.

3. The sentence completion, comprehension, and production materials were developed as a means of investigating patterns of language deficiency in educationally

subnormal children of primary school age.

4. These tests were also being used as tests of developmental language proficiency with younger normal children as well as with groups of Asian and West Indian immigrant children.

AN INFORMAL LANGUAGE INVENTORY

John D. Melear Colorado
1973

Purpose

This study was concerned with the lack of testing materials which could be used to measure a child's spontaneous language. Melear felt that freedom of expression is essential in any attempt to estimate the quantity and quality of a child's real oral language abilities and thus adapted work done in the area to construct such a measurement procedure.

Methodology and instrumentation

A group of 338 elementary children from Weld County, Colorado, with a mean chronological age of 7.6 years was selected. All of these students lacked basic vocabulary skills as measured by the Peabody Picture Vocabulary Test. A similarly aged control group with normal basic vocabulary skills was also selected.

Both groups had the spontaneity and correctness of their oral speech measured according to the following procedure. A paraprofessional supplied a group of children with art supplies and encouraged them each to draw a picture (ten minutes). Next he turned on a tape and asked, "Who would like to tell us about his picture?" The paraprofessional waited only forty-five seconds for a response. In those situations where no response occurred, he went on to other activities. However, the response rate was high (program group response rate was seventy-seven percent, control group response rate was eighty-four percent) and child after child chose to describe his art work. All recordings were transcribed and evaluated by one person. This procedure was titled the Informal Language Inventory.

Analysis

The analysis of the spontaneously collected student responses considered number of grammatically correct sentences, percentage of totally correct sentences and mean morphemes per respondent. Tables showing these results were illustrated.

Outcomes and interpretations

1. The major deficiencies of the program group children were in the areas of vocabulary and grammatical structure.

2. These results confirm past studies.

3. Further analysis of the instrument could be made to

include counts of descriptive terms and modifying phrases.

4. The Informal Language Inventory appears to have promise in measuring aspects of the oral language program not covered on current commercial tests.

5. This technique is convenient to administer since a group of examinees can be involved, with scoring of their responses done at a later time.

TESTING COMMUNICATION

Adrian S. Palmer Michigan
1972

Purpose

The purpose of this study was to answer two questions. First, was there an oral communication ability not being measured, even indirectly, by a usual foreign language proficiency test battery? Second, could this ability be objectively and reliably measured?

Methodology and instrumentation

Two types of experimental tests were constructed.

1. COMTEST was intended to measure two-way oral communication ability. The subject was shown a set of pictures or drawings. The examiner told him that he was thinking about one of the set and the subject must ask questions until he could identify which picture it was. COMTEST contained two examples and ten problems. The number of pictures or drawings in each set ranged from three to seven.

2. PROTEST was intended to measure oral production ability. The subject's task was to describe orally one out of four similar pictures so that the examiner could determine which of the pictures was being described. The test contained one example and ten items.

3. Additional control tasks
   a. ACTEST was an experimental test of aural comprehension in which the subject was shown two pictures. A recorded narration contained four clues which identified one of the two pictures. The subject was scored on the time it took him to correctly identify the correct picture.
   b. Michigan Proficiency Test Battery (MPT)
      i. Michigan Test of Aural Comprehension (MAC) consisted of ninety recorded problems with written answer choices.
      ii. Composition (COMP) was a thirty-minute impromptu composition written on an assigned topic.
      iii. Michigan Test of English Language Proficiency (MT) was a 100-item objective test on grammar, vocabulary, and reading comprehension.

   PROF was the mean of scores on the three subtests.

All tests were administered to (1) non-native speakers who were

enrolled in the English Language Institute, (2) non-native speakers enrolled as regular students at the University of Michigan, and (3) native speakers of English.

Analysis

1. Means, standard deviations of the experimental tests were calculated.

2. Experimental test scores and ACTEST scores were converted to normalized T-scores.

3. These normalized T-scores were compared with scores on the MPT battery.

4. Multiple correlation coefficients were calculated and the Spearman Brown formula was applied.

5. A factor analysis was performed on the transformed scores of the experimental tests, MPT battery scores, an index of sex and the index of status.

Major outcomes and interpretations

The experimental tests measured factors not being measured in the traditional language proficiency test battery such as oral communication proficiency.

A MEANS OF ACCOUNTABILITY IN ORAL COMMUNICATION PERFORMANCE

Robert E. Potter
David B. Strother

Washington
1975

Purpose

The authors developed and reported on an accountable procedure for measuring and recording counter-productive verbal and nonverbal responses presented by a student in an oral production.

Methodology and instrumentation

The article contained two illustrations from classroom practise of the authors' method of recording precise verbal and nonverbal responses at work. In both cases, the procedure followed was to (1) select one specific behavior, either desirable or undesirable; (2) count the number of recurrences of that specific behavior within a certain period of time; (3) determine the recurrence rate per minute by dividing the total number of responses by the total number of minutes in the observation period, and (4) plot the behavior on a sic cycle graph. This provided an objective base rate for the student to decrease in the case of an undesirable response or to increase in the case of a desirable response. In further oral presentations the student was signaled by the teacher whenever the response was made. A typical case cited resulted in the rate of occurrence dropping from 16.2 per minute to 2.2 per minute for an inappropriate verbalization. Data was collected in several classes as part of regular instruction in oral production.

Analysis

Data was collected and analyzied to illustrate the effectiveness of the procedure. The data consisted of tabulations and graphs. The logical and valid interpretation of this data was suggested to be criterion-referenced in that a student's response pattern of concern was compared to a suitable standard, and work was started to reach the standard.

Major outcomes and interpretations

1. An interesting procedure was illustrated for objectively recording student response patterns in oral productions.

2. The procedure has promise for objective continuous measurement of both student progress in oral language and the effectiveness of a speech development program.

ORAL EXPRESSION TESTS

Peter Robinson Quebec 1969
Toronto 1971

Purpose

The purpose of this set of papers was to describe problems and possible solutions to oral language testing and reading comprehension testing whether the language was a first or second language. Advantages and disadvantages of differing amounts and kinds of student participation in the testing tasks were discussed.

Methodology and instrumentation

For the purposes of his work, Robinson divided student test participation into four categories for each of oral language testing and reading comprehension testing. This summary will be concerned with the oral language testing only, where the four levels of participation were zero, limited, extended, and complete. Examples are:

1. Zero Participation
   The student hears a group of sentences. He is asked a question. He selects what he believes to be the most appropriate one from a certain number of answers.

2. Limited Participation
   The student hears a group of sentences. He is asked a question. He gives his own answer.

3. Extended Participation
   The student hears a group of sentences. He is asked to make a summary.

4. Complete Participation
   The student hears a group of sentences. He is asked to make a comment.

Robinson's work stressed the use of student participating modes for valid testing. He discussed numerous examples of participating test situations.

Analysis

Samples of participating test situations were analyzed in the paper. Suggestions for rating the speaking ability of the examinee were given.

Major outcomes and interpretations

Robinson believed the test designer should create a situation which would encourage the subject to give maximum expression because his language in the test is a representative sample of what he can produce in a real situation. The study recommended the use of an interview and discussion section in any oral language test. The author pointed out that there are no widely accepted linguistic criteria for the evaluation of oral expression but that the two major concerns should be comprehensibility and acceptability. Robinson suggested a standardized record should accompany the tapes of all oral tests when they are being used to assess oral competency.

The record should include duration of test, number of words produced by the subject, number of sentences, type of sentences, type of grammatical structures, type of lexical structures, type of phonetic structures, type of grammatical errors, type of lexical errors, and type of phonetic errors.

THE DEVELOPMENT OF A MEASURE TO EVALUATE LANGUAGE COMMUNICATION SKILLS OF YOUNG CHILDREN

Suzanna Rose Jim Maxwell Pennsylvania
Margaret Wang Elaine Corey 1973

Purpose

The purpose of this study was to develop an evaluation instrument to assess young children's language communication skills, specifically to assess the quality of inter-communication among young children.

Methodology and instrumentation

Two parallel sets of tasks were developed in the Language Communication Skills Task (LCST). It was designed for two children playing two different roles of a communication task, to jointly solve an inter-communication problem - successful placement of an object on the receiver's picture board based on the verbal message transmitted by the presenter. The presenter's board contained all of the objects in the appropriate place on the board while the objects for the receiver's board were in an array on the table. The players were not allowed to look at each other's pictures nor use hand gestures. They could use language as much as they needed prior to the placement of the object by the receiver. The LCST was designed to assess language communication competencies of both the receiver and presenter.

Subjects included in the study were randomly selected from an inner city public elementary school in Pittsburgh. There were thirty-eight kindergarten students, thirty-two grade ones and forty-two grade twos. Mean I.Q.'s as measured by the Slossen Intelligence Test were 102.47, 102.33, and 93.86 respectively. Children were randomly paired within a grade level and each child played both roles in different sessions. There was no time limit although the mean time was twenty-five minutes. Each session was taped and location of objects on the receiver's board was recorded.

Analysis

1. Communication Measures

   The following four measures were used:
   a. presenter's score;
   b. receiver's score,
   c. pair score represented by correct placement of the object according to the direction given; and
   d. inter-communication score represented by correct placement of the object according to the predetermined

location which resulted from the use of the encoding and decoding skills of both the presenter and receiver.

2. Linguistic Measures

These measures included:
a. tokens--number of words used;
b. type--total number of different words used;
c. token length--mean number of letters per word for the total words used;
d. type length;
e. type-token-ratio--a measure of variability;
f. a measure of repeativeness; and
g. utterance length--number of words included in a meaningful unit of verbalization preceeded and followed by a pause.

The linguistic measures were derived directly from computer analysis of the prototypes (Maxwell 1973).

3. Reliability of LCST was estimated by test-retest, split-half, and parallel test methods.

4. Correlations were calculated between the two parallel tests, between subject's receiver and presenter scores, between order of presentation and subscores and among numerous combinations of variables.

5. Analysis of variance between the inter-communication means of each age group was performed.

6. Item analyses were carried out for the LCST.

Major outcomes and interpretations

1. This study indicated that the LCST is a reliable and valid technique to use for assessing particular communication skills.

2. An increase with age in the communication proficiency and linguistic proficiency was observed.

3. The LCST can serve as a diagnostic technique for identifying characteristics of competencies that lead to adequate and effective communication.

MEASURING ORAL PRODUCTION IN ENGLISH AS A FOREIGN LANGUAGE

B. H. Seward

Egypt
1973

Purpose

The purpose of this study was to determine whether the results of objective paper and pencil tests of general language proficiency could serve as measurements of students' speaking proficiency in English as a Foreign Language (EFL).

Methodology and instrumentation

Ninety-four out of one hundred eighty students were selected as subjects from classes of the English Language Institute of the American University in Cairo (ELI). For seventy-eight of the ninety-four subjects, scores from the Michigan Test of English Language Proficiency (MICH) and a composite ELI course score were obtained. Two teachers from each class were asked to rate the subject's speaking ability in English. These results were averaged to give a single rating.

Analysis

The ratings were correlated with total and part scores on the MICH.

Major outcomes and interpretations

The study concluded that scores from objective paper and pencil test batteries indirectly provide much more effective and stable measurements of students' speaking proficiency than teacher evaluations.

180

THE CONSTRUCTION OF AN ORAL COMPOSITION QUALITY SCALE

J. Skull
A. Wilkinson

Great Britain
1969

Purpose

The purpose of this research was to produce a quality assessment scale for one kind of spoken English at the 15+ to 16+ level. This assessment scale would use a semantic differential to achieve its purpose.

Methodology and instrumentation

The construction of the scale involved administering an oral question to a selection sample of thirty-nine students from a larger sample of 130 final year CSE students. The students spoke individually for three to four minutes on the selected topic, "One of the Best Times in My Life." These talks were graded by four experienced teachers of English, and from the pooled judgment of these assessors, ten talks were selected that seemed to represent the widest spread of ability. These ten talks were again taped in random order.

A new group of forty assessors were asked to mark the ten talks. Half the group marked by general impression (score out of ten) and half used a semantic differential (score out of seven) on twelve pairs of terms.

Analysis

The semantic differential scores were totalled and then expressed out of ten, so comparison with the general impression scores was possible. Further analysis was computed on the scores arrived at both ways. Finally, from the grand mean scores on the master score sheet (that is, the pooled judgment of forty assessors) an oral composition scale was produced by putting subjects in rank order.

Major outcomes and interpretations

1. A scale for one particular kind of oral composition (a personal account, usually involving narration and description) was constructed.

2. It is considered that this scale could be used by teachers to estimate their pupils' oral composition of this particular kind. They could do this by comparing a pupil's oral production with the samples on the scale and then assigning the scale-value of the sample that

most nearly equals their pupil's oral production.

3. The semantic differential was not found to be an effective measure of oral production in this study. However, further work in the area may correct the weaknesses found.

4. An interesting result of this study was the perfect correlation of the scores on the significant linguistic features with the ranked positions on the composition scale.

A COMPARISON AND MEASUREMENT OF ORAL LANGUAGE GROWTH

Russell G. Stauffer — Delaware
John J. Pikulski — 1974

Purpose

Stauffer and Pikulski's study was concerned with measuring the growth in oral language skills of grade one children evidenced by dictated stories. The recorded dictated stories served as an assessment of a particular reading program being offered in their district.

Methodology and instrumentation

The analysis of oral language was made on the basis of materials obtained in the notebooks of fity children who were selected randomly from a larger group of 232 children. Both pre-test and post-test information was collected. The types of dictation collected were group-dictated stories and individual-dictated stories. Also, the project involved the collection of new words in word banks which were used in both oral language work and creative writing.

Analysis

Analysis of the dictated stories involved the variables of average number of words per dictation, average number of sentences, number of words in longest sentence in a selection, average number of pronouns, average number of prepositions, and average number of different words. Results from the selected dictations were tabulated under the above headings.

Outcomes and interpretations

1. Students used both basic sentences and longer sentences reflecting the use of modifiers, clauses, and other elaborations.

2. A wide variety of topics were dealt with at the grade one level. In fact, a total of 3331 different words were tabulated.

3. Results showed that when stories dictated by a group of first grade children were analyzed, significant improvements were found along all dimensions of oral language usage being evaluated. This supported the language experience approach in language arts as an instructional procedure.

4. The research indicates that this is a very effective way of assessing oral language growth where oral language development is stressed as part of the instructional program.

TESTING AND EVALUATIVE PROCEDURES IN SPEECH COMMUNICATION

Charles O. Tucker Illinois
1970

Purpose

The purpose of this article was to review issues concerning speaker evaluation. In this article, speech communication was defined as being formal and informal, public and private, individual, group and mass communication. It also included co-curricular programs in debate, individual events and oral interpretation.

Major outcomes and interpretations

1. A major part of the evaluation process in speech involves efforts to focus attention on the student in a way in which he has not seen himself or others before.

2. Speech behavior is changed by changes in the standards of evaluation. Speech evaluation and speech instruction proceed together.

3. The most extensively used and varied form of assessment of single speeches is oral and/or written criticism. This is followed by the use of a rating scale.

4. Audience response to a single speaker can be measured by comprehension tests and semantic differentials designed to do that job. Very little attention has been given to the testing of delayed message effect.

5. Tests such as Rokearch's open-and-closed-mindedness test can be used to evaluate individuals as speakers. Other dimensions, using different tests, can also be measured.

6. Individual assessment can occur at three stages - during speech preparation, during speech presentation and afterwards in the form of analysis. Assessment should involve self-assessment and listener assessment.

7. Small group communication assessment involves problems in both measuring many separate speech units and considering the group qualities that may be influencing the communication.

8. Factors that should be considered in small group assessment are group productivity (actual communication), group climate, and leadership structure.

AN ANALYSIS OF ORAL LANGUAGE COMPARED WITH READING ACHIEVEMENT

Maurice Williams
Edith Weinstein
Ralph O. Blackwood

1970

Purpose

The purpose of this study was to determine oral language of elementary school children and compare levels between grades.

Methodology and instrumentation

Subjects were twenty students from each of grades one, three, and five from five schools. Subjects were of low average socioeconomic background, had average I.Q. and academic ability and were considered non-failures.

Three pictures from the Chandler Language Arts Series were shown to each pupil during the seventh month of the school year. Each pupil was asked three series of three questions designed to initiate sequentially a literal, an interpretive, and a critical level of thought responses. The oral reactions were tape recorded and transcribed. Using the Botel Readability Formula, the average grade level of each pupil's oral language was determined.

Analysis

The Median Test was applied separately to both boys' and girls' scores across the three grade levels.

Major outcomes and interpretations

1. The results indicated that girls made no gains in level of conversation during the first five years of school. Boys gained only between the first and third grades.

2. It would appear that present teaching methods for this sample were of little or no consequence in developing oral language achievement.

THE QUALITY OF LANGUAGE EXPERIENCES IN YOUNGER CHILDREN

Andrew Wilkinson Great Britain
1969

Purpose

The research reported concerned the construction of tests of "oracy," both of listening and speaking, in the secondary school area. In addition, useful information was collected on the nature of the spoken language and the kind of criteria by which we should be able to judge it.

Methodology and instrumentation

The paper described numerous case studies which illustrate different factors in language development. A comparison of language development in a compulsory situation and non-compulsory situation was given.

Major outcomes and interpretations

1. The author's main concern was not with having children talk more but with improving the quality of talk as they progress through school.

2. Listings of locations where work in oral language is progressing were given.

APPENDIX E:

SUMMARIES OF RESEARCH RELATED TO ASSESSMENT OF LISTENING

ORAL LANGUAGE: LISTENING
CHAPTER 10 OF A LANGUAGE FOR LIFE

Sir Allan Bullock Great Britain
1975

Purpose

The purpose of this section was to describe the present situation in England with respect to the teaching and assessing of listening skills.

Report content

This report contended that the difficulty of listening is commonly underestimated, and is highly related to human attention, that is, the longer people are required to listen, the less effectively they do so. Research has indicated that of the time devoted by students to listening, speaking, reading, and writing, well over half is taken up by listening. Further, when listeners had the opportunity to reply or participate through action, their attention was stimulated, and they listened better.

The author questioned whether present techniques for improving listening skills produce either short or long term gains. Further, it was questioned if there was any relation between listening test scores obtained in an artificial test situation and in actual listening behavior. The administration of "listening exercises" to a whole class was criticized, and the suggestion made that a more individualized approach was essential.

Major Outcomes and interpretations

1. The authors suggested that listening ability was part of a highly complex process related to the individual situation, and to the knowledge and experience of the listener, the nature of his motivation, and the degree of his involvement.

2. It was also contended that a conscious policy on the part of the teacher is essential to extending student's receptive capacities.

COMPARISON OF LISTENING ABILITIES, CATEGORIZED AS GOOD AND POOR, OF INNER-CITY CHILDREN IN THE SIXTH GRADE

Joseph Butler

Indiana
1970

Purpose

The purpose of this study was to compare listening abilities of inner-city children in the sixth grade.

Methodology and instrumentation

This study involved 124 sixth grade students. For each student, scores were available for reading, intelligence, and school achievement. The STEP Listening Test was administered to each subject.

Analysis

Listening Test scores were compared to reading scores, intelligence scores, and achievement level of each subject.

Major outcomes and interpretations

1. Students classified as good listeners were found to have significantly higher scores on tests of word meaning, paragraph meaning, and intelligence.

2. The STEP Listening Test appeared to have merit as an instrument to assess listening ability of inner-city children.

3. The ability to listen well has positive implications for scholastic achievement.

LEARNING FROM PROSE MATERIALS

Ronald P. Carver District of Columbia
1973

Purpose

The purpose of this research was to study the relationships among presentation rate, understanding, information stored, and learning of prose passages to see if different empirical and conceptual types of measurements would produce the same results as in previous studies.

Methodology and instrumentation

The 108 subjects for this study were recruited through advertising among the student body at the University of Maryland. The prose passages used were taken from the Carver-Darby Chunked Reading Test (Carver & Darby, 1970). The subjects were divided into ten groups, half of which listened to the prose passages at varying rates of presentation (seventy-five wpm to 450 wpm) the other half reading the passages with varying time limits.

Subject understanding of the prose was empirically indicated by subjective ratings, that is, subjects were asked to estimate the percentage of the total number of complete thoughts in a passage that they understood. Stored information was indicated by a chunked type of test which was purported to validly indicate this variable. Learning was indicated by two types of cloze procedures, one a twenty per cent cloze test where one out of every five words was deleted, the second a revised cloze test in which two consecutive words out of every five words were deleted.

Analysis

Mean scores were calculated for each group by rate of presentation and by test. Correlations between understanding means and test score means were also calculated.

Major outcomes and interpretations

1. The results indicated that for both the listening subjects and reading subjects, one of the most reliable and easiest methods of investigating the understanding of prose materials was simply to ask the subjects to estimate the percentage of thoughts that they understood.

2. The data suggested that the cloze method is relatively invalid as a measure of what is understood in reading. This was based on the fact that while understanding scores dropped eighty per cent as rate of presentation increased, cloze scores dropped only twenty per cent.

3. It was further found that chunked test scores were almost perfectly related to understanding. On a percentage scale the two scores varied in an almost one-to-one relationship. This data supported an earlier hypothesis of Carver's (1971) that the degree of understanding of prose and the degree of stored information is one and the same.

A DESCRIPTIVE STUDY OF THE GOLDMAN-FRISTOE-WOODCOCK TEST OF AUDITORY DISCRIMINATION AND SELECTED READING VARIABLES FOR PRIMARY SCHOOL CHILDREN

Ronald L. Finkenbinder Pennsylvania
1973

Purpose

The purpose of this study was to explore the reliability and validity of the Goldman-Fristoe-Woodcock Test (G-F-W Test) and its relation to selected reading variables.

Methodology and instrumentation

Two hundred forty-two students from kindergarten through grade three were randomly selected (sixty per grade) from two elementary schools. In January, 1971, the Metropolitan Readiness Test (MRT) was administered to the kindergarten sample, and the Gates MacGinitie, Primary A, B, and C were administered respectively to the grade one, two and three samples. The Wide Range Achievement Reading Test, and the GFW were individually administered to all the subjects. The GFW was readministered in May, 1971, to forty-nine subjects randomly selected from the total group.

Analysis

Correlation coefficients were used to analyze results. Only a reliability coefficient of .80 or higher was considered to be of substantial magnitude to facilitate interpretation of the GFW.

Major outcomes and interpretations

1. For the GFW Quiet subtest, this study found that most children began to have complete success (thirty out of thirty correct) by grade two, thus substantiating the test's original design to test strengths rather than weaknesses. This characteristic probably would identify children with auditory discrimination problems more readily after grade two.

2. The GFW Noise subtest was more difficult than the quiet subtest for all groups, but only significantly so for the kindergarten group. The author questioned the usefulness of this subtest as a predictor of auditory discrimination problems.

3. In the readministration of the GFW, significant improvement on both subtests was found. However, test-retest reliability for this group of children on the two subtests was not high (.62, .42). Internal consistency correlations carried out for each subtest at the four grade levels revealed only one substantial correlation (Grade three, Quiet subtest).

4. Neither of the GFW subtests correlated highly with either readiness or reading variables. Use of this instrument as a predictor of reading ability for these grade levels is questionable.

A NEW DIMENSION IN THE ASSESSMENT OF SPEECH SOUND DISCRIMINATION

Ronald Goldman
Macalyne W. Fristoe
Richard W. Woodcock

Tennessee
1971

Purpose

After perusal of the available tests, the authors decided there was a need for a test to evaluate speech-sound discrimination ability while minimizing the influence of variables such as abstractness of materials, or familiarity with vocabulary and illustrations used. This new test was primarily designed for preschool and primary age children.

Methodology and instrumentation

The Goldman-Fristoe-Woodcock Test (G-F-W Test) was a pictorial test designed to provide measures of speech-sound discrimination ability which were relatively unconfounded by other factors. It provided a measure of auditory discrimination under ideal listening conditions plus a comparative measure of auditory discrimination in the presence of controlled background noise.

In designing the test, the effects of extraneous factors on auditory discrimination were minimized in the following ways:

1. all test words were chosen from among those familiar to young children,
2. a prerecorded tape was used for presentation of all test items,
3. subjects utilized both immediate and past experience (short term memory and long term memory) rather than the traditional approach to discrimination testing based primarily on short-term memory.

The test was comprised of three parts, a Training Procedure, a Quiet Subtest, and a Noise Subtest. In all parts, the subject was asked to point to a picture of a word presented to him as a stimulus.

Analysis

The design of the test provided a method of assessing the subjects performance on different dichotomous sound categories so that the examiner could determine if the subject had unusual difficulty with a particular category. The manual provided separate norms for each subtest from ages three years, eight months to seventy

years and over, as well as providing a comparison of children in different diagnostic categories (ie learning disabilities, delayed speech, mental retardation, etc.).

Major outcomes and interpretations

1. Further research may show this test useful in devising therapy plans for specific types of individuals.

2. The provision of two scores (ideal conditions versus background noise conditions) may provide a means for locating individuals with deficits in auditory attention.

A HIERARCHY OF SKILLS IN LISTENING COMPREHENSION AND READING COMPREHENSION

Marie G. Hackett Florida
1970

Purpose

This study was concerned with the applicability of a hierarchically ordered achievement test to the diagnosis and assessment of listening and reading skills.

Methodology and instrumentation

Performance objectives were derived from eleven language comprehension skills and arranged in a hierarchical structure. Two parallel items were constructed to yield pass-fail information for each of the eleven skills. The test was administered to 1,186 subjects at the second, fifth, eighth, and eleventh grades. The test consisted of four levels A, B, C, and D, for grades two, five, eight and eleven respectively. Subjects were randomly assigned to one of two testing conditions, oral condition or reading condition. The performance on each skill was marked pass if both items were correct, and fail if one or both items were incorrect.

Analysis

At each grade level, conditional probabilities relating performance on lower-level skills to higher level skills were calculated.

Major outcomes and interpretations

1. With respect to sequence and transfer of skills, conditional probabilities of correct responses to listening measures indicated patterns of predictable relationships of lower-level to higher-level skills at all four grade levels studied.

2. The author suggested that two questions per skill area, while enabling the collection of informative data, were insufficient to draw significant conclusions.

VISUAL AND AUDITORY TRAINING WITH THE COZE PROCEDURE TO IMPROVE READING AND LISTENING COMPREHENSION

D. K. Kennedy
Paul Weener

Pennsylvania
1973

Purpose

The purpose of this study was to determine the effectiveness of individualized training with the cloze procedure to improve reading and listening comprehension.

Methodology and instrumentation

Four groups of twenty third graders who were below average in reading were used. Two experimental groups were trained individually with the cloze procedure, one group with a visual mode, the second group with an auditory mode of presentation. Of the two control groups, one group received an individualized oral reading program while the other group remained in the regular classroom. The effects of the experimental program were assessed using the Durrell Listening-Reading Test, and two cloze procedure post-tests, one presented orally, the other visually.

Analysis

The group grade score means and standard deviations for the two subtest scores of the Durrell Listening Reading Test were obtained. Similarly, means and standard deviations for number of errors on the listening and reading post tests of the cloze procedure were obtained. For both measures, four by two analysis of variance with repeated measures was done to determine if the observed differences in the means were significant. Tables showing these results were illustrated.

Major outcomes and interpretations

1. The experimental reading group did significantly better than any of the other three groups on the Durrell reading comprehension subtest. The indication was that visual training with the cloze procedure improved reading comprehension.

2. The experimental listening group scored higher than the control class group, but not significantly higher than the control reading group or the experimental reading group on the Durrell listening comprehension subtest.

3. The experimental listening group did no better on the Durrell reading comprehension subtest than the two control groups indicating no significant transfer from auditory training to the reading comprehension test.

4. Both the visual and auditory training procedures facilitated the ability to perform on the post-test cloze procedure. This may have resulted from a shift in the attentional processes during reading from word units to units larger than a word, and hence increased ability to use contextual cues.

5. A final implication is that visual training in cloze procedures, which takes no special training, and which can be applied to reading materials already in use, should be considered in remediation programs for below-average readers.

MEASUREMENT OF VERBAL LISTENING ACCURACY IN CHILDREN

Merlin J. Mecham Utah
1971

Purpose

The purpose of this paper was to describe the development of a testing technique to measure verbal listening accuracy in children that would provide reliable measures from group as well as individual administrations.

Methodology and instrumentation

The author suggested that many children who are poor readers are good listeners, and that such children may be predestined to fail because they were obliged to use a poorly functioning channel of communication when a satisfactorily functioning channel was available.

An eighty-six item test was developed, requiring a word-picture matching response on the part of the child who heard a label and identified its matching picture accurately. The test could be either individually administered, or administered to a group using a tape recorder to present stimulus words.

The test was administered on a classroom group basis to 1,857 elementary school age children from grade two through six in the Salt Lake City area. All subjects were considered to be normal achievers. An additional nineteen children were tested individually.

Analysis

Internal consistency of the test was used to estimate its reliability (KR_{20}). Mean test scores by age and by sex were also calculated.

Major outcomes and interpretations

1. The internal consistency reliability of the test was within acceptable limits for all levels (.859-.982). Individual administration did not increase the reliability estimate.

2. The author concluded that listening accuracy as measured by this test appeared to be a developmental phenomenon which plateaus rapidly between the 4th and 6th grades.

AUDING ACHIEVEMENT OF FIRST GRADE PUPILS RELATED TO SELECTED PUPIL CHARACTERISTICS

Mary C. Nesbitt Georgia
1968

Purpose

This study was designed to identify the auding levels of achievement of first grade children, and to compare the auding achievement of children in the first grade who had one year of pre-primary instruction with children in the first grade as their first year of schooling.

Methodology and instrumentation

An auding instrument was constructed using four selections from Laidlaw Brothers Science Series for elementary grades. Questions required answers about main ideas, details, organizations, and inference. A tape-recorder was used to present the selections to each of eighty first graders.

Analysis

Results were statistically treated to determine if a significant difference in auding achievement existed between the two groups. The proportion of students able to comprehend materials designed for higher grade levels was also determined.

Major outcomes and interpretations

1. No significant difference between the auding achievement of the two groups was found.

2. Through listening fifty-six per cent of the subjects could understand a selection designed to be used at the second grade level.

3. Through listening thirty per cent of the subjects were able to understand a selection designed to be read at the fifth grade level.

SOURCE OF GAINS IN LISTENING SKILL: EXPERIMENTAL OR PRE-TEST EXPERIENCE

Robert A. Palmatier
George McNinch

New York
1972

Purpose

The purpose of this study was to determine if growth in listening ability could be produced as an incidental effect of a planned program in an area such as the study skill of note-taking.

Methodology and instrumentation

Subjects for this study were 135 eleventh grade college preparatory students enrolled in a parochial school in Syracuse, New York. The average I.Q. for the group was one standard deviation above the national norm. The total sample was divided into four treatment groups.

Pre-testing of all subjects included the administration of the Watson-Glazer Critical Thinking Appraisal (Form Ym), Spitzer Study Skills Test, Taped Note Taking Test (Form P) and the Brown-Carlson Listening Comprehension Test (Form AM). Scores from nationally administered achievement tests were also collected. The four treatment groups were then exposed to an eleven-day schedule of testing and training in note-taking skills, presented primarily via taped lessons. Post-testing included the administration of the Brown-Carlson Listening Comprehension Test (Form Bm), and the Taped Note-Taking Test (Form c).

Analysis

Analysis of coveriance using the pre-test score on the Brown-Carlson as the controlling criterion was undertaken to determine if significant gains had been achieved. Correlation analysis was carried out to determine if a relationship existed between gains in listening skill and measured achievement scores.

Major outcomes and interpretations

1. When pretest scores on the Brown-Carlson were used as the coveriant to negate the learning effects and influence of the pretest administration on the post-test administration, gain in listening skill was not significant.

2. Post-test scores on the listening test correlated very highly with the pretest scores on the alternate form.

3. Correlation analysis revealed negligible relationships between gain in listening skills and standardized test achievement, and gain in listening skills and intelligence. The author hypothesized from this that listening skill is a discrete entity not easily predicted by other measurable educational attributes.

4. The author concluded that the Brown-Carlson Test of Listening Comprehension did not have sufficient equality of forms nor test-retest reliability to support its use as a pre- and post-treatment measure.

LEARNING TO LISTEN: AN EXPERIMENTAL STUDY

D. A. Penfield
L. A. Marascuilo

California
1972

Purpose

This study was aimed at determining whether or not it is possible to train students to become better listeners relative to a given set of skills, and to determine whether there were socioeconomic or sex differentials in the acquisition of these skills.

Methodology and instrumentation

A test based on eleven basic skills involved in effective listening and reading was developed by Hackett. This test consisted of three questions assessing each of the eleven skills, and was used by the authors to measure their subjects' listening skills.

The subjects included fifteen grade two classes and nineteen grade five classes in Alameda County, California. The classes at each grade level were randomly selected to either training conditions (experimental) or no training conditions (control). Students in the experimental condition were given training in the eleven skills.

Analysis

Three variables were controlled in the analysis: socioeconomic status (SES), sex, and training. Mean scores were computed for each of the resulting populations.

Major outcomes and interpretations

1. For grade two students, scores in the high and low SES categories were not significantly different nor were those between boys and girls. However, scores were significantly higher for those who received training than for those in the no training condition indicating that training was effective for second grade students.

2. At the grade five level, although the scores favor training, only the differences between high SES boys and girls and low SES boys were found to be significant. Thus, by grade five, there appeared to be a distinct change in performance favoring students at the high SES level. The implication was that environmental conditions play a stronger and stronger role in the acquisition of information via listening as students progress through school.

3. The results indicated that training in listening skills narrows the gap between SES levels, particularly at the higher grades.

THE INFLUENCE OF THREE LISTENING-ENVIRONMENT FACTORS ON LISTENING COMPREHENSION OF FOURTH-GRADE STUDENTS

Jack Albert Price

Texas
1969

Purpose

This study was designed to determine the effect of purpose, background information and distractions on listening skills.

Methodology and instrumentation

Subjects for this research consisted of 692 fourth grade students from twenty-four classrooms. Classes were randomly assigned to eight treatment conditions. All classes were exposed to a twelve minute taped dramatization, however, half the classes were told they would be tested on the taped material while the other half were given no purpose for listening. Two classes from each group were exposed to background material about the taped story prior to listening to it. One half of the classes in each group were subjected to distractions on the tape, and to common aural and visual distractions while listening to the tape while the other classes were kept as free from distractions as possible. All subjects were asked to respond to a twenty-item multiple choice test designed to measure the subjects' comprehension of the story content.

Analysis

Results for all eight treatment conditions were tested for differences.

Major outcomes and interpretations

1. Having a purpose for listening did not significantly increase listening efficiency.

2. Having prior background information did not significantly affect listening comprehension.

3. Listening efficiency was significantly decreased by distractions in the listening environment.

THE AUDITORY ANALYSIS TEST: AN INITIAL REPORT

Jerome Rosner
Dorothea P. Simon

Pennsylvania
1971

Purpose

The purpose of this study was to develop an individually administered auditory perception test for elementary children which would determine

1. if test results indicated a range of individual competencies in the behaviors defined as auditory perception;
2. if individual test results correlated with measures of reading performance at different grade levels; and
3. if item analysis of results could provide input to the design of training programs to teach competent auditory perceptual skills.

Methodology and Instrumentation

The Auditory Analysis Test (AAT) was designed to measure whether a child had acquired the skills needed for competent analysis and synthesis of the phonemes presented in a beginning reading program. It consisted of forty English words varying in length from one to four syllables. Each word was pronounced by the examiner who instructed the child to repeat the word, then to repeat it again but to omit a specified sound. The sounds to be eliminated included thirteen different consonants, one consonant digraph, two consonant blends, and twelve syllables. Seven different word types were included.

The test was administered individually to 284 children in kindergarten through grade six in a middle class suburban elementary school in Western Pennsylvania. For each administration, testing was discontinued after four consecutive errors.

Analysis

Mean scores and standard deviations were determined for each of the seven grade levels administered the AAT. Individual AAT scores were correlated with the sum of the stanines on the language arts subtests of the Stanford Achievement Test, as well as with Otis Lennon I.Q. scores. Item difficulties were calculated by grade level for each question.

Major outcomes and interpretations

1. Results showed that the ability to respond appropriately to the AAT improved for all seven word types in step with maturation and/or improved reading skills for children from kindergarten through third grade. Only the three more difficult word types reflected changes related to age and grade within grades four, five and six.

2. The major differences evidenced between kindergarten and grade one may imply that reading instruction is an important factor in teaching a child to perform the tasks measured by the AAT.

3. A substantial correlation was found between AAT and language arts scores. The possibility exists, therefore, that an intervention program designed to improve AAT performance may improve reading achievement.

4. The item analysis results provided information for improving the validity and reliability of the AAT. It was hypothesized that these results could also provide input for the structuring of a training program.

5. AAT could well be used to help assess the effectiveness of new instructional programs designed to assist the child in developing more analytical auditory skills.

AUDITORY ANALYSIS TRAINING WITH PREREADERS

Jerome Rosner Pennsylvania
1974

Purpose

The purpose of this study was to determine the extent to which auditory analysis skills could be taught to four-year-old children.

Methodology and Instrumentation

The author defined the term "auditory analysis" as the resolution of spoken words into their phonomic elements. The goal of auditory analysis training then is to teach the child a process for identifying the acoustical elements of the reading-spelling code as heard in the context of spoken language.

The subjects used for this study included twenty-six inner-city preschoolers, sixty-two kindergarten children from the same inner-city school and fifty-seven kindergarten students from a suburban school in Pittsburgh.

The auditory analysis testing and training program used for the study was taken from the Perceptual Skills Curriculum (Rosner, 1972), a program comprised of sequential instructional goals with a criterion-referenced test for each objective. Before commencing the training program, each child was pretested (T-1) to determine the number of objectives he had already mastered. The inner-city preschoolers and kindergarten suburban students were exposed to the program for nine months. At that time final testing (T-2) was undertaken on all three groups.

Analysis

The auditory analysis skills of the inner-city preschool group at the completion of training was compared to the skill level of both the inner-city and suburban kindergarten control groups at the beginning of the training program.

The mean number of objectives mastered by each of the three groups for (T-1) and (T-2) and statistical differences between those means were calculated. A chart of the results was included.

Major outcomes and interpretations

1. The data clearly indicated that for the children included in the study, auditory analysis skills could be taught to prereaders. Mean (T-2) scores for the preschoolers were significantly higher than mean (T-1) scores of either the inner-city or suburban kindergarten groups.

2. The results supported the argument that it is feasible to teach auditory analysis skills well in advance of reading instruction, thus dealing effectively with at least one aptitude that is closely related to reading achievement.

VARIABLES INFLUENCING PERFORMANCE ON SPEECH SOUND DISCRIMINATION TESTS

Arthur H. Schwartz
Ronald Goldman

Tennessee
1974

Purpose

It has been hypothesized that factors in the construction and administration of speech-sound discrimination tests can influence performance. The purpose of this study was to investigate the effect of variations in the context of stimulus-item presentation on speech-sound discrimination under different background listening conditions.

Methodology and instrumentation

Seventy-two preschool, kindergarten, and first grade children were divided into three groups on the basis of age. All subjects had been screened on several criteria before inclusion as subjects. Stimulus items consisted of thirty monosyllabic nouns that were visually depicted in line drawings. These drawings were divided into ten sets, each of which contained three members that differed from each other only by their initial sound. The stimulus items were presented in three different contexts:

1. a paired-comparison context in which only a pair of words were presented;
2. a carrier-phrase context in which the same sentence was used to present each stimulus item;
3. a sentence context in which four different sentences were alternated to present stimulus items.

Each stimulus context was presented as a noise subtest and as a quiet subtest.

Analysis

Performance was measured by counting the number of error responses for each of the six subtests. A four-factor analysis of variance was used to determine the effects of order, age, context of stimulus-item presentation, and background conditions on speech-sound discrimination.

Outcomes and interpretations

1. At each age level, significantly more errors were made in the paired-comparison context than in carrier phrase or sentence context. This suggested that some aspect of this manner of assessing speech-sound discrimination adversely influenced performance.

2. Differences between the carrier phrase and sentence-context were not significant indicating that these two contexts may be highly similar in nature.

3. For each stimulus context, a decrease in the number of error responses occured with succeeding increases in age level.

4. Under noise conditions, all age groups made more errors, however, noise appeared to have the most adverse effect on performance of the paired-comparison context.

5. The higher number of errors obtained from the paired-comparison context suggested that in speech-sound discrimination tasks, children rely on linguistic as well as acoustic cues.

THE DEVELOPMENT OF A LISTENING COMPREHENSION TEST FOR KINDERGARTEN AND BEGINNING FIRST GRADE

Nancy K. Wallner Mississippi
1974

Purpose

The purposes of this study were to develop a test of listening comprehension for beginning first grade, and to establish the reliability and validity of the instrument.

Methodology and instrumentation

Using the Spache Readability Formula and the Dale-Chall Readability Formula, twelve passages were written, ranging in difficulty from grades one to four. The twelve passages were paired into two equivalent forms (A and B) with respect to levels of difficulty. Fourteen multiple-choice questions were written for each passage, seven measuring literal comprehension and seven measuring inferential comprehension. The three possible alternatives for each question were in the form of pictures. All instructions, passages, questions and possible responses were recorded on tapes.

Both forms of the Wallner Test of Listening Comprehension (TLC) were administered to 150 kindergarten students in Farmington, Missouri. Each student was supplied with a headset, and was separated from other subjects by cardboard screens.

Analysis

Item-analysis was completed to determine how the test discriminated among subjects. Descriptive statistics were computed for both forms, and a t-test applied to compute the significance of differences between means and between variances. Total scores for each of the two forms of the TLC were correlated with one another, and with subtest scores of the Metropolitan Readiness Tests (MRT) to establish alternate form reliability and criterion-related validity. To establish predictive validity, total scores for both forms of TLC were correlated with subtest scores of the Stanford Achievement Test (SAT) administered a year later.

Major outcomes and interpretations

1. The two forms of the TLC could not be considered parallel since their means were not statistically equal. However, neither the variances or K-R_{20} reliabilities were found to be significantly different.

2. The alternate-form reliability of the TLC, KR_{20}=.892, was considered as very high.

3. Both forms of the TLC were shown to have sufficient predictive validity with respect to reading achievement to warrant their use for educational purposes.

A HIERARCHY OF LISTENING SKILLS

Susan W. Weaver
William L. Rutherford

Texas
1974

Purpose

The purpose of this report was to provide a hierarchy of listening skills stated in terms that would permit both teachers and parents to evaluate easily what the child can and cannot do.

Methodology and instrumentation

In the absence of research to indicate a specific sequence for auditory abilities, the author derived a hierarchy from current programs and theories of development of auditory abilities drawn from a review of literature. The hierarchy was composed of three sections: environmental skills, discrimination skills, and comprehension skills. The skills listed were concentrated in infancy, pre-school, and primary grades, however the author indicated that the skills would broaden and extend through life. Certain skills were also listed that would apply only to visually handicapped persons.

Major outcomes and interpretations

1. This hierarchy could be used by both teachers and parents to assess what a child can and cannot do, and from an initial evaluation determine the skills which might be expected to develop.

2. The hierarchy could be used in remediation cases where some higher order skills were learned before some more basic skills, but where the deficit in the basic skills would impede continued development of the more advanced listening skills.

THE QUALITY OF LISTENING

Andrew Wilkinson
Leslie Stratta
Peter Dudley

Great Britain
1974

Purpose

One of the purposes of this study, which commenced in 1966, was to construct measures of listening appropriate to the age range eleven to eighteen.

Methodology and instrumentation

One of the major aims in the construction of the three test batteries was to avoid the use of written English read aloud, and to concentrate on using material that people are hearing naturally most of the time, that is, the spoken language. Wherever possible, natural conversation that had been recorded on tape served as a basis for test passages.

In total, three batteries, each consisting of a number of subtests, were constructed. Each subtest concentrated on one of the following elements of listening comprehension: content, contextual constraint, phonology, register, or relationship. Once the initial material for each battery was prepared, all the questions were field tested in a number of schools in the Birmingham area. Following the analysis of the piloted materials, each battery was revised as necessary. In their final form, Battery A was intended for use with ten to eleven year olds, Battery B with thirteen to fourteen year olds and Battery C with seventeen to eighteen year olds.

Analysis

The revised batteries were administered to extensive samples of students thus making it possible to develop a scale of normalized and standardized scores. Reliability for each battery was determined using a split-half method. Finally, test results from each listening Battery were correlated with other standardized test scores.

Major outcomes and interpretations

1. The reliabilities obtained for Batteries A, B and C respectively were 0.827, 0.777, and 0.838.

2. Significant positive correlations were found between scores on the listening tests (Batteries A and B) when compared to both a reading test and an intelligence test. Significant positive correlations were also found between the total listening scores on Battery C and the scores on the Senior English test and an intelligence test.

3. The authors stressed that the emphasis in their listening tests was on the richness and variety of language rather than upon "skills". As such, they would see at least part of the test's purpose as directed towards developing an awareness of the features and functions of spoken language.

SELF EVALUATION IN SPEAKING AND LISTENING

George L. Williams New York
1972

Purpose

The purpose of this report was to describe the use of Listening and Speaking Self-Evaluation charts to help students examine what they hear and say.

Methodology and instrumentation

The author pointed out that students should be encouraged to evaluate for themselves by developing criteria to make judgments about a speaker, book, or report. In a New York public school system, English teachers had been preparing a series of questions for the self-evaluation of listening and speaking. An example of the kinds of questions included at the grades four to six level were:

1. After hearing a paragraph, can you identify an omitted or misplaced sentence in the paragraph?
2. After hearing a paragraph in which a word is omitted, can you supply an appropriate missing word?
3. After listening to a paragraph, can you identify the transitional words in it?

Major outcomes and interpretations

1. The use of listening charts could help a student who is unaware of his listening potential to become cognizant of what to listen for.

2. The author suggested that use of self-evaluation charts could help to develop the critical thinking skills of the student.